THE BLOOMSBURY COMPANION TO PHILOSOPHY OF PSYCHIATRY



EDITED BY
ŞERIFE TEKIN AND
ROBYN BLUHM

BLOOMSBURY

The Bloomsbury Companion to Philosophy of Psychiatry

The Bloomsbury Companion to Philosophy of Psychiatry

Edited by
Şerife Tekin and Robyn Bluhm

BLOOMSBURY ACADEMIC
LONDON • NEW YORK • OXFORD • NEW DELHI • SYDNEY

BLOOMSBURY ACADEMIC
Bloomsbury Publishing Plc
50 Bedford Square, London, WC1B 3DP, UK
1385 Broadway, New York, NY 10018, USA

BLOOMSBURY, BLOOMSBURY ACADEMIC and the
Diana logo are trademarks of Bloomsbury Publishing Plc

First published in Great Britain 2019

Copyright © Şerife Tekin and Robyn Bluhm, 2019

Şerife Tekin and Robyn Bluhm have asserted their rights under the Copyright, Designs and Patents Act, 1988, to be identified as Editors of this work.

For legal purposes the Acknowledgments on p. xv
constitute an extension of this copyright page.

Series design by Eleanor Rose
Cover image © Pichet Ounta / Alamy Stock Photo

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage or retrieval system, without prior permission in writing from the publishers.

Bloomsbury Publishing Plc does not have any control over, or responsibility for, any third-party websites referred to or in this book. All internet addresses given in this book were correct at the time of going to press. The author and publisher regret any inconvenience caused if addresses have changed or sites have ceased to exist, but can accept no responsibility for any such changes.

A catalogue record for this book is available from the British Library.

A catalog record for this book is available from the Library of Congress.

ISBN: HB: 978-1-3500-2408-3
ePDF: 978-1-3500-2406-9
eBook: 978-1-3500-2407-6

Series: Bloomsbury Companions

Typeset by Integra Software Services Pvt. Ltd.

To find out more about our authors and books visit www.bloomsbury.com
and sign up for our newsletters.

Contents

Notes on Contributors viii
Acknowledgments xv

Part 1 Introduction 1

1 Introduction to Philosophy of Psychiatry *Şerife Tekin and Robyn Bluhm* 3

2 What Is Psychiatry? *Mona Gupta* 19

Part 2 Philosophy of Mind and Psychiatry 39

3 Understanding the Nature of Mental States: Psychiatry, the Mind-Body Problem, and the Biopsychosocial Model of Medicine *Jesse Butler* 41

4 "Aren't Mental Disorders Just Chemical Imbalances?," "Aren't Mental Disorders Just Brain Dysfunctions?," and Other Frequently Asked Questions about Mental Disorders *Ginger A. Hoffman* 59

5 Psychopathy, Autism, and Basic Moral Emotions: Evidence for Sentimentalist Constructivism *Erick Jose Ramirez* 93

6 Anorexia Nervosa and the Embodied Mind *Lana Kühle* 113

Part 3 Phenomenology and Psychiatry 131

7 Merleau-Ponty and the Foundations of Psychopathology *Anthony Vincent Fernandez* 133

8 Transformation through Dialogue: Gadamer and the Phenomenology of Impaired Intersubjectivity in Depression *Constantin-Alexander Mehmel* 155

Part 4 Philosophy of Science and Psychiatry 175

9 Philosophy of Science, Psychiatric Classification, and the *DSM* *Jonathan Y. Tsou* 177

10 Inductive Risks and Psychiatric Classification *Aaron Kostko* 197

11 Causal Explanation in Psychiatry *Tuomas K. Pernu* 217

12 Trauma-Informed Psychological Research *Ami Harbin* 237

Part 5 Ethics and Psychiatry 255

13 Informed Consent in Psychiatry: Philosophical and Legal Issues *Claire Pouncey and Jon F. Merz* 257

14 The Ethics of Coercion and Other Forms of Influence *Kelso Cratsley* 283

15 Voice, Silencing, and Listening Well: Socially Located Patients, Oppressive Structures, and an Invitation to Shift the Epistemic Terrain *Nancy Nyquist Potter* 305

16 Mental Illness, Agency, and Responsibility *Michelle Ciurria* 325

17 Philosophers, Psychopaths, and Neuroethics *Matthew Ruble* 347

Part 6 Social and Political Philosophy and Psychiatry 369

18 Neurodiversity Theory and Its Discontents: Autism, Schizophrenia, and the Social Model of Disability *Robert Chapman* 371

19 Managing Individuals and Populations through Psychiatric Classification *Devonya Havis and Melissa Mosko* 391

20 The North African Syndrome: Traversing the Distance to the Cultural "Other" *Bryan Mukandi* 413

Part 7 Metaphysics, Epistemology, and Psychiatry 431

21 Mental Disorder, Free Will, and Personal Autonomy *Christian Perring* 433

22 Metaphysical Problems in Psychiatric Classification and Nosology *Peter Zachar* 447

23 Scientific Realism, Antirealism, and Psychiatric Diagnosis
Sam Fellowes 467

24 Social Epistemology and Psychiatry *Anke Bueter* 485

Part 8 Philosophy of Medicine and Psychiatry 505

25 The Functions of Diagnoses in Medicine and Psychiatry
Hane Htut Maung 507

26 Mental Health and Illness: Past Debates and Future Directions
Reinier Schuur 527

Annotated Bibliography/Further Reading 543
Index 554

Notes on Contributors

Aaron Kostko is Assistant Professor of Philosophy in the Center for Learning Innovation at the University of Minnesota Rochester. His research focuses on methodological, conceptual, and ethical issues that arise in neuroscience and psychiatric research, classification, and treatment. He also focuses on epistemological and ethical issues that arise in evidence-based methods to research effective teaching and student learning.

Ami Harbin is Associate Professor of Philosophy and Women and Gender Studies at Oakland University. She is the author of *Disorientation and Moral Life* (Oxford University Press, 2016). Her research focuses on moral psychology, feminist ethics, and bioethics.

Anke Bueter holds a PhD in philosophy from Bielefeld University and currently works as Assistant Professor at the Leibniz University Hanover, Germany. Her work focuses on the intersection of social epistemology and philosophy of science, especially philosophy of medicine and psychiatry. Her publications span topics such as the impact of feminism on medicine, the role of values in science, and epistemic injustice in psychiatric classification.

Anthony Vincent Fernandez is Assistant Professor of Philosophy at Kent State University in Ohio. He works on classical and contemporary phenomenology, focusing especially on the challenges of applying phenomenology to the study of psychopathology. He has recently published in *Journal of Consciousness Studies*, *Phenomenology and the Cognitive Sciences*, *Synthese*, and *Philosophy, Psychiatry, & Psychology*. He is also coeditor of *The Oxford Handbook of Phenomenological Psychopathology*.

Bryan Mukandi studied medicine at the University of Zimbabwe and interned at the United Bulawayo Hospitals. He subsequently pursued postgraduate studies in the social sciences and humanities at the National University of Ireland, Galway and Queen's University of Belfast. Having worked in health systems and policy research for some years, he completed his PhD in Philosophy at the University of Queensland, in Australia, with a thesis exploring what Steve Biko referred to as "Black Consciousness." Bryan has published work in the *South African Journal of Philosophy*, *Australian Journal of Indigenous Education*, and

the *American Journal of Public Health*. He is currently Lecturer in Medical Ethics in the University of Queensland's Faculty of Medicine.

Christian Perring is Adjunct Full Professor at St. John's University, Queens, New York. He is currently Vice President of the Association for the Advancement of Philosophy and Psychiatry, and his areas of research focus are the moral responsibility and autonomy of people with mental illness, the psychiatric treatment of young people, and personal identity. He has been editor of Metapsychology Online Reviews for over twenty years. His enthusiasms include indie music, craft beer, and hiking.

Claire Pouncey, MD, PhD, is a psychiatrist in private practice in Philadelphia. She has worked in many psychiatric settings, including psychosomatic medicine. She uses her formal training in philosophy and medical ethics to write about psychiatric epistemology and ethics, topics that include nosology, psychiatric recovery, professionalism, informed consent, and the Goldwater Rule.

Constantin-Alexander Mehmel is a PhD student in Philosophy at the University of Essex. His PhD is funded by the Consortium for the Arts and Humanities South-East England (CHASE) Doctoral Training Programme, and he is working toward a phenomenological account of the experience of disorientation. Prior to starting his PhD, he was Visiting Researcher at the Heidelberg University Hospital working on the intersection of hermeneutics and social cognition, particularly enactivist approaches to intersubjectivity. His research interests are primarily in phenomenology, philosophy of emotion, philosophy of psychiatry, social cognition, and hermeneutics (esp. Gadamer).

Devonya Havis is Associate Professor of Philosophy at Canisius College in Buffalo, NY. Her scholarly engagements utilize insights from Michel Foucault as a means of exploring issues in Critical Philosophy of Race, Critical Disability Studies, and Phenomenology. Havis's teaching and research are motivated by a keen interest the ways philosophy can be used to prompt ethical interventions that contest injustice. Her publications include "'Now, How You Sound': Considering a Different Philosophical Praxis" in *Hypatia*; "The Parrhesiastic Enterprise of Black Philosophy" in *Black Scholar*; and she is a contributor to *Addressing Ableism: Philosophical Questions via Disability Studies*. She has also been featured on the "Dialogues on Disability" and "PhilosopHER" blogs.

Erick Jose Ramirez received his doctoral degree from the University of California at San Diego and is currently Assistant Professor of Philosophy at Santa Clara University. His research centers on issues relating empathy, moral agency,

mental illness, and the philosophy of technology. Specifically, he is interested in drawing interdisciplinary connections between empirical research on emotion, empathy, and mental illness and philosophical theories of value, agency, and philosophical methodology. The latter, he argues, can benefit from incorporating more of the former. His most recent research investigates the degree to which virtual reality technologies can be used to improve philosophical methodologies and enhance our empathic perspective-taking abilities. He also investigates how VR creates new ethical issues for researchers and game designers. His work has most recently appeared in *Philosophical Psychology*, *Metaphilosophy*, *Ethics in Information Technology*, *Philosophical Explorations*, and the *American Journal of Bioethics: Neuroscience*.

Ginger A. Hoffman is Associate Professor of Philosophy at Saint Joseph's University. She possesses backgrounds in both philosophy (her PhD in philosophy is from MIT) and neuroscience (her PhD in neuroscience is from Yale University, where she researched the basic biology underlying antidepressant and antipsychotic mechanisms). Her current work is in the philosophy of psychiatry, philosophy of disability, philosophy of mind, value theory, and feminist philosophy. More specifically, she employs advances in these areas of philosophy to the mission of making psychiatric and neuroscientific practices more ethical and more rigorous. She studies, among other things, the impact of relational and "4E" models of the mind and self on psychiatric diagnosis and treatment, possible defenses of Mad Pride and neurodiversity paradigms, and sanist/anti-sanist or feminist/anti-feminist implications of neuroscientific and psychiatric research. She also works with mentally divergent individuals in a volunteer context and engages in various forms of mental diversity advocacy.

Hane Htut Maung is Leverhulme Research Fellow in Philosophy at the University of Manchester. His research focuses on the conceptual, metaphysical, and ethical problems in medicine and psychiatry. He also has broader interests in the philosophy of mind and the philosophy of science. Prior to his career in philosophy, he read medicine and specialized as a psychiatrist in the UK's National Health Service. He holds a PhD in philosophy from Lancaster University, an MPhil in philosophy and ethics of mental health from the University of Pretoria, and an MB BChir in medicine from the University of Cambridge. He is a member of the Royal College of Psychiatrists.

Jesse Butler is Associate Professor in the Department of Philosophy and Religion at the University of Central Arkansas, where he teaches courses in the philosophy of mind, critical thinking, philosophy of language, and philosophy

for living. His research focuses on theories of self-knowledge, with his book *Rethinking Introspection* (Palgrave Macmillan, 2013) offering a pluralist and naturalistic approach to understanding the first person perspective. Dr. Butler was also a US Fulbright Scholar during the 2016–17 academic year, teaching courses on cross-cultural approaches to self-knowledge at Jinan University in Guangzhou, China. His current research involves the application of theories of self and mind to issues in the philosophy of psychiatry and environmental philosophy, toward the end of cultivating human well-being and sustainability through understanding the interconnected plurality of our embodied existence in the world.

Jonathan Y. Tsou is Associate Professor of Philosophy at Iowa State University. He has published philosophy of psychiatry journal articles and book chapters on topics such as natural kinds in psychiatry, the neurobiological basis of mental disorders, the social feedback effects ("looping effects") of psychiatric classifications, and the classification of mental disorders by the *DSM*.

Jon F. Merz is Associate Professor in the Department of Medical Ethics and Health Policy in the Perelman School of Medicine at the University of Pennsylvania. Dr. Merz teaches Research Ethics and Bioethics and the Law in the Department's Masters of Bioethics curriculum. His research interests encompass privacy and confidentiality in medicine and research, reproductive rights and policy, research ethics and regulation, conflicts of interest, informed consent, intellectual property, and issues raised in the foregoing areas by biotechnology. Dr. Merz's current activities include funded research and training programs in Tanzania and Guatemala. Since early 2003, Dr. Merz has been the moderator of the IRB Forum [http://www.irbforum.org], a discussion forum and resource for IRB professionals across the globe.

Kelso Cratsley is currently a Visiting Researcher at the Kennedy Institute of Ethics, Georgetown University. He has previously taught at George Washington University, University of Massachusetts Boston, and Concordia University in Montreal, and was recently Visiting Scholar at the University of California, San Diego. His research centers on issues in bioethics, philosophy of psychology and psychiatry, and global health policy.

Lana Kühle is Assistant Professor of Philosophy at Illinois State University. She received her PhD from the University of Toronto. Her areas of expertise are in philosophy of mind and phenomenology with close interdisciplinary connections to neuroscience and psychology. Her overall goal is to understand

how one's bodily awareness gives rise to the sense of self and the subjectivity of conscious experience. She has published several articles including "The Subjectivity of Experiential Consciousness: It's Real and It's Bodily" in *Mind & Matter* and "The Missing Pieces in the Scientific Study of Bodily Awareness" in *Philosophical Psychology.*

Matthew Ruble is Visiting Assistant Professor of Philosophy in the Department of Philosophy & Religion at Appalachian State University in Boone, North Carolina. He earned a BA in Philosophy & Religion and an MA in Psychological Counseling, both from Appalachian State University. After several years working in community mental health services Ruble returned to school earning an MA in the Philosophy and Ethics of Mental Health from Warwick University, UK. Following years of grant writing and administering educational opportunity programs for low-income students, Ruble returned to school once more to earn his PhD in Philosophy from the University of Tennessee. (It's never too late!) His dissertation, *Responsibility, Blame and the Psychopath,* was directed by Adam Cureton and Professor Emeritus John Hardwig and focused on conceptions of mental illness that create serious problems for our understanding of moral responsibility. When he is not doing philosophy Ruble can be spotted climbing and running in the beautiful Blue Ridge Mountains with his dog, Ella.

Melissa Mosko is Associate Professor of Philosophy at Canisius College in Buffalo, NY. Her teaching focuses on the intersections of feminist theory and political theory; she is especially interested in the effects of social power on individual moral and political agency. She has published on philosophy of psychiatry, existentialism, political theory, social epistemology, and business ethics. Her current research is on debates about the legalization and decriminalization of sex work.

Michelle Ciurria is Visiting Scholar and Lecturer at the University of Missouri-St. Louis. She previously held postdoctoral fellowships at the University of New South Wales (2016–17) and Washington University in St. Louis (2014–16). She completed her PhD at York University in Toronto in 2014. Dr. Ciurria's work explores the intersections between moral responsibility, moral psychology, social epistemology, critical race theory, and feminist sociology. She is developing a relational account of moral responsibility that takes social injustice as the starting point for theorizing about moral responsibility, identifying inequalities in our responsibility system, and developing fairer responsibility practices. Dr. Ciurria's work has appeared in such venues as the *Journal of the American Philosophical Association, Philosophical Psychology, Feminist Philosophy Quarterly*, and *Amy*

Schumer and Philosophy. In her spare time, Dr. Ciurria enjoys political activism, volunteering in her community, blogging, playing with her cats, and spending time with her family.

Mona Gupta, MD CM, FRCPC, PhD, is a psychiatrist and researcher in bioethics at the Centre l'Hospitalier de l'Université de Montréal. Her broad area of academic interest is the intersection of ethics and epistemology in psychiatry including topics such as the nature of evidence in clinical practice. Her research monograph on ethics and evidence-based psychiatry was published by Oxford University Press in 2014. Her current work focuses on how psychiatrists identify and make use of relational and testimonial evidence in clinical reasoning.

Nancy Nyquist Potter is Professor of Philosophy at the University of Louisville. She has published extensively in feminist ethics and political philosophy as well as the interdisciplinary field of philosophy and psychiatry. Her current research is on the connections between theories of knowledge and ethics, with special attention to nosological and diagnostic issues, and on the relationship between voice, silence, and uptake, in particular for patients living with mental illness. Her last book-length publication was a critical analysis of borderline personality disorder (Oxford University Press, 2009) and her newest book, *The Virtue of Defiance and Psychiatric Engagement*, is now out (Oxford University Press). She is a Core Faculty Member of the Interdisciplinary Master's Program in Bioethics and Medical Humanities, an Adjunct Faculty with Women and Gender Studies, and an Adjunct with the Department of Psychiatry and Behavioral Sciences.

Peter Zachar, PhD, is Professor of Psychology at Auburn University Montgomery. He was the chair of the psychology department for eight years, is currently the Associate Dean in the College Sciences, and the President of the Association for the Advancement of Philosophy and Psychiatry. Zachar is the author of over 100 publications, including *Psychological Concepts and Biological Psychiatry: A Philosophical Analysis* (John Benjamins, 2000), and *A Metaphysics of Psychopathology* (MIT Press, 2014).

Reinier Schuur is a PhD Candidate in the Department of Philosophy at the University of Birmingham.

Robert Chapman is currently a Teaching Associate in Philosophy at the University of Bristol. He recently completed a PhD on the ethics of autism at the University of Essex, and has previously taught at both the University of Essex and at King's College London. He is neurodivergent, and works at the intersection between the philosophy of psychiatry, bioethics, and critical disability studies.

Sam Fellowes is philosopher of psychiatry and historian of autism. His philosophical interests primarily focus upon the metaphysical and epistemological status of psychiatric diagnosis, specifically, what are they and how much do we know of them. His historical interests primarily relate to comparing older notions of autism with modern notions to establish the level of similarity and to see if older notions contain important insights relevant for potential modifications to current notions. Since completing his PhD at Lancaster University, he has started a Wellcome Trust Research Fellowship. This project is entitled "Neo-Kantianism as an alternative conceptual basis for psychiatry" and is also at Lancaster. He is primarily interested in the history and philosophy of psychiatry because he is himself autistic. He wishes to gain a greater understanding of how adequate current and potential alternative notions of autism are for both scientific and practical purposes.

Tuomas K. Pernu, PhD, is Research Fellow at King's College London (Department of Philosophy) and University of Helsinki (Molecular and Integrative Biosciences Research Programme, Faculty of Biological and Environmental Sciences). He is Research Associate in Philosophy & Medicine at King's (King's College London), an initiative that actively brings together philosophers and medical practitioners. Dr. Pernu's research interests revolve around the topic of causation and causal explanation, including fundamental theoretical issues in metaphysics of causation as well as more pragmatic issues related to philosophy of science and scientific explanation. He is also a cycling entrepreneur, a farmer, an Ordained Minister of The Church of the Latter-Day Dude, and a fan of Phil Collins.

Acknowledgments

The editors would like to thank Rachel McKenzie for her invaluable assistance with manuscript preparation. Members of the Executive Council of the Association for the Advancement of Philosophy and Psychiatry provided great help with the early stages of planning this volume and also useful feedback on the introduction and the plan for the book; we would especially like to acknowledge comments from Jennifer Radden and from Claire Pouncey. Finally, we are indebted to the editorial assistance of Cortez, Clementine, and Gus.

Part One

Introduction

1

Introduction to Philosophy of Psychiatry

Şerife Tekin and Robyn Bluhm

Introduction

You are taking a philosophy class in the second semester of your sophomore year at college. The course requires close reading of fairly difficult historical texts, based on which the professor leads lively lectures and debates in class. Your friend April complains to you that she doesn't understand anything in class. You advise her that she make sure to do the readings line by line, by paying attention to detail, and staying focused on the discussion that takes place in class, as you observe that she is rather distracted. She keeps checking her phone every few minutes, not listening to the professor during the lectures. She never participates in class discussion. You can tell that the professor and other students in class are irritated by her late arrival to and disruption of class. No one wants to do group homework with her especially after she lost one of your classmate's lecture notes. It was rather surprising for you to see April like this because she took a couple of classes with you during her freshman year and she had performed well.

A couple of months later April leaves class just as you are about to write a test. You wonder why she is leaving. Next class you ask her what happened, and she tells you that she will be writing her tests in a quiet room by herself in the library. After asking you promise not to share it with anyone, she explains why. Her parents took her to a psychiatrist at the advice of few of her professors who had noticed the changes in April's performance and that the psychiatrist diagnosed her with this illness called attention deficit and hyperactivity disorder (ADHD). It turns out that was why she was as distracted and inattentive in class. April doesn't want her other friends know because she is worried that they will make fun of her.

You are rather puzzled: as your little brother, too, is diagnosed with ADHD, but his behavior does not at all resemble April's. Your brother is super active,

running around at inappropriate times and places, and talks a lot, even when it is not his turn to speak. He was so disruptive in classes, displaying erratic behavior, that his elementary school teacher contacted your parents and asked them to take your brother to a psychiatrist. You are surprised that both April and your brother have the same diagnosis because they behave completely differently: April is quiet, distracted, and inattentive, while your brother is loud, fidgety, intense, and active. You are also puzzled to witness the change in your feelings about April: initially you had thought that she was being a bad student who doesn't respect her classmates or professor and disrupts class, but now you feel bad about having thought that way because it was the illness causing her to behave this way.

You also start to wonder how many of the apparently disruptive people who are not very good at what they do actually suffer from a mental disorder. Then you start asking yourself how is it that we distinguish a student who is unmotivated, or just a poor student, from one who is suffering from a mental disorder. If the same illness can express itself in completely different behavior how exactly can doctors detect it? Is April responsible for her behavior in class, or did her illness cause her problems? What can be done to help students like April so that they are not ashamed of their mental illness? How do you prevent others making fun of people with mental disorders? You may not realize it, but when you start asking these questions, you are doing philosophy of psychiatry.

Philosophy of psychiatry is a branch of philosophy that will help you answer, or at least make sense of, some of these questions pertaining to the experience of mental disorder, its scientific research and clinical treatment, as well as various social and political questions it raises. To understand how philosophy of psychiatry will guide you through these questions, we first review what philosophy is and then explain how the fundamental problems of psychiatry connect to the main branches of philosophy.

What is philosophy?

Philosophy, according to its Greek etymology, means "love of wisdom." Beyond that, however, it becomes difficult to define, in part because there are a number of different, and sometimes conflicting, approaches to "love," "wisdom," and thus, "philosophy." One approach to defining philosophy is to look at its methods and goals, as the philosopher Bertrand Russell does. Russell suggests that philosophy results from a critical examination of the grounds of our convictions, prejudices,

and beliefs. It unifies and systematizes the body of the sciences, for Russell (1912). Thus, philosophy uses systematic and critical reflection, to search for truth exempt from prejudices, misconceptions, and falsehoods, and thereby to expand our conceptual framework for developing a true understanding of the world.

Another way to understand what philosophy is to think about its different branches, each of which aims to answer different kinds of questions. In what follows, we take this approach, which allows us to also show how the different branches of philosophy are related to philosophical questions about psychiatry. Specifically, we review philosophy of mind, philosophy of science, ethics, and social and political philosophy.

Philosophy of mind and psychiatry

There is a close link between philosophy and psychiatry. As discussed in detail by Mona Gupta in Chapter 1, psychiatry is the branch of medicine that aims to understand the causes of mental disorders, and to develop scientific intervention techniques to help those who suffer from them. One of the fundamental branches of philosophy is philosophy of mind, and it concerns the connection between the mind and the body by answering questions such as the following: How does the mind connect to the body? In other words, how do our mental states, such as beliefs and desires, or psychological states such as anger, relate to our physical states, such as the physiological processes of the brain or hormonal mechanisms? The answer to this question is intimately relevant to the problems of psychiatry; that is, if we understand how mental states relate to physical states, we can also understand what is going on in the body when, say, a person suffers from mental states such as depressed mood. For example, if you understand how mental states relate to physical states, you will be able to explain how your brother's or your friend April's ADHD causes different behaviors in each of them. Or, you would be able to make sense of how medications might regulate their behavior.

Additional questions in the philosophy of mind literature that are relevant to psychiatry include what the self is, that is, what makes someone the person they are. For example, you've heard your aunt say that your brother has become a different person since he started taking medication for his ADHD. You understand what she means, but have always thought it sounded strange to say so. Obviously, he is still *himself*, but what *makes* him that person? Would he still

be the same person if he didn't have ADHD, or is his ADHD part of what makes him himself? If ADHD is part of who she is, did April become a different "self" in your sophomore year from the one she was in freshman year?

For an accurate investigation of mental disorders, we not only need information provided from a third person perspective on mental disorders through scientific research, but also need an understanding of the first person encounter with mental disorders. The best resource for this is patients' examinations and descriptions of their experiences of mental disorders. Phenomenology, as a philosophical study of human experience, can also be considered as a branch of, or approach to, philosophy of mind. It offers resources for philosophers of psychiatry to contemplate the "what it is like"-ness of mental disorders. Phenomenologists investigate the essential features of consciousness of human subjects, including selfhood, intersubjectivity, affectivity, and temporality. They explore how a meaningful, lived world is opened up for us, and how we make sense of it. In this respect, one of the best ways for you to understand what it is like to live with ADHD may be to talk to your friend April and your brother about their experiences with the condition, how they perceive their illness, as well as how the diagnosis of illness affects their lives.

Philosophy of science and psychiatry

Philosophy of science asks questions about the nature of things like scientific theories and explanations, as well as the nature of science itself. Increasingly, philosophers of science are also concerned with how best to use the results of scientific investigation to effect change in the world. Psychiatry, and medicine more broadly, is related to science, because scientific knowledge is necessary to medicine. And psychiatric research, for example in neuroscience and in psychology, is clearly an area of science.

It is commonly believed that the central aim of science is to develop an accurate picture of the causal structure of the world. If science helps us to understand causal relationships between phenomena, then we can use this knowledge to act effectively. For example, if we predict the speed of hurricane Irma as well as the locations it is anticipated to affect, we can take precautions so as to prevent damage to human lives. For example, we can encourage individuals who live in the most dangerous regions to evacuate their homes. Psychiatry, as a health science, has similar goals; it aims to offer an understanding of the causes and mechanisms of mental disorders in order to facilitate effective clinical and

preventative interventions. One way in which it works toward these goals is to identify the generalizable features of mental disorders. For example, if we learn, through scientific research, what the features of depression are (such as significant changes in individuals' sleeping habits, or weight, or significant loss of interest in previously enjoyed activities), we would be more successful in identifying a depressed person and therefore helping them receive treatment. Thus, the way in which we classify mental disorders, as say whether someone with psychological distress should rightly receive a diagnosis of ADHD, is crucial for psychiatrists to develop the right strategies for intervention. Going back to the thought experiment above, for example, psychiatrists identify two different types of ADHD, which is the reason for April and your brother having the same diagnosis but displaying different behaviors. The one that April suffers from is called inattentive type, while your brother's is called the hyperactive-impulsive type.

But given our discussion of philosophy of mind, it's natural to wonder how such different kinds of behavior are related to activity in the brain. If mental disorders are caused by changes in brain activity, does this mean that both April and your brother have the same underlying brain changes? A major goal of psychiatric research is to answer that question—and there is a lot of debate among both psychiatrists and philosophers of psychiatry over whether the current diagnostic categories used in psychiatry (as outlined in, for example, the *Diagnostic and Statistical Manual of Mental Disorders [DSM],* which, developed by the American Psychiatric Association, offers the standard criteria for the classification of mental disorder and used widely in North America and around the world, and the *International Classification of Diseases and Related Health problems [ICD],* which, developed by the World Health Organization, is the international "standard diagnostic tool for epidemiology, health management and clinical purposes") do reflect a specific physical process. Psychiatry is unique in medicine in that diagnoses are based on psychological and behavioral criteria, rather than on a physical examination and the kinds of laboratory tests used in other medical specialties. There is also a great deal of debate over whether this is a good thing. Some people argue that understanding the biology of mental disorders is the best way for us to develop more effective treatments—and that this might require that we get rid of current diagnostic categories and replace them with more accurate ones. Others argue that focusing on biology will lead us to lose sight of the fundamentally psychosocial nature of mental disorders. How we should classify and explain ADHD and other mental health issues is very much an open question.

Another scientific goal in psychiatry is to depict the underlying causes of mental disorders. For example, if we know the underlying causes of ADHD are, we may develop better strategies to treat it so that people like your friend April or your brother do not suffer. The power of scientific research in psychiatry is that we can develop generalizations that will guide clinicians not only to help one or two individuals inflicted with mental disorders, but develop a systematic understanding, and therefore interventions for a whole population of those with mental illness. If we categorize similar phenomena under the same category, then what we know about one individual under that category may also be valid for another individual in the same category.

Questions about the nature of scientific classification and scientific explanation are central to philosophy of science and these have bearings on the science of psychiatry. Philosophers of science, for example, deliberate on whether there is a single, fundamental level of explanation to which all other explanations reduce. Reductionists accept this claim whereas pluralists maintain that explanations should involve multiple disciplines and span multiple levels. Connectedly, a fundamental debate in philosophy of psychiatry is whether mental disorders can be individuated and explained as a disorder of the brain, that can be delineated as the neuroscience matures (reductionism), or whether mental disorders can and must be explained at the multiple levels, not just the cellular level, but at the individual, social, and cultural levels (pluralism), requiring the cooperation of and collaboration with multiple sciences. Reductionists may argue, for example, that April and your brother's ADHD will eventually be explained as an anomaly in their brain mechanisms, while pluralists would argue that we need multiple scientific approaches to understand ADHD not just neuroscience.

Ethics and psychiatry

Philosophers of psychiatry also discuss ethical issues related to how we should perceive and treat people who have been diagnosed with a mental disorder. For example, April received accommodations to write her exams after being diagnosed with ADHD. Some of your classmates or you may have thought initially that it is not fair for her to receive this type of special support to do well in her exam, however you may now understand that, from an ethical standpoint, the special features of her illness (e.g., being easily distractible) warrant her receiving support to be able to write her exam in a quiet room by herself. Also,

remember how your attitude toward April has changed once you heard that she suffers from ADHD; instead of feeling frustrated with her inattentiveness in class you started feeling sympathy for her. But many people who have been diagnosed with a mental disorder experience stigma: people view them more poorly and treat them more harshly because of their diagnosis.

Another important theme in psychiatry is the capacity of those with serious mental disorders to make rational decisions and their ability to act as rational, autonomous agents, problems that are at the heart of ethics. If someone with a mental disorder lacks rational capacity and ability to act autonomously, then it is not clear whether the person bears the full responsibility for their actions in the spaces of morality and the law. Determination of the agency and autonomy of persons with mental disorders has profound implications for their clinical treatment. Some key questions span the nature of informed consent in psychiatry, the ethics of forced treatment, and insanity defense in the law.

Social and political philosophy and psychiatry

There are many questions about psychiatry and about mental disorders more generally that are related to issues in social and political philosophy. Social and political philosophy aims to understand how the social and political systems influence individuals. Key questions in philosophy of psychiatry pertaining to the themes of social and political philosophy include the influence of medical, social, and political systems on the emergence and treatment of mental disorders; connections between oppression and mental health; the relationships among sexism, racism, and other forms of bigotry and how they are supported by the institutional structures and systems listed above. We are all profoundly shaped by our culture, so understanding the social and cultural make-up of a society is important to understanding the experiences of, and attitudes toward, individuals with mental disorders. For example, it may be interesting to explore whether the social response to April and your brother would be different based on their gender.

A second kind of social and political question related to philosophy of psychiatry is how we should understand the behaviors and experiences that we have categorized as mental disorders. The idea that psychiatry is a branch of medicine and, therefore, is best approached from a scientific perspective, has become dominant in clinical practice and medical education, as well as research and research funding. However, important concerns about the limits of this

perspective are still being raised. Here we describe two related kinds of challenge to the medical model in psychiatry that can be situated under the social and political philosophy.

During the 1960s, a controversial "anti-psychiatry" movement became prominent. At this time, a number of people, among them practicing psychiatrists, began to argue that the experiences and problems we associate with such conditions as depression and schizophrenia are not actually signs of illness. Rather they are reactions to personal problems and/or to intolerable circumstances. Thomas Szasz, who famously described mental illness as a myth, argued that we should instead understand people's experiences and behaviors as reactions to "problems in living" (Szasz, 1961). He believed that, if something is a disease, then it involves a malfunction of a physiological process, and that behaviors therefore cannot be diseases. Moreover, because the conditions it "treats" are *not* diseases, psychiatry should not provide medical (e.g., pharmacological or surgical) treatments. This is especially important in cases where an individual was facing the prospect of being treated against their will (because they had been deemed to be incompetent to refuse treatment) or could not give valid consent to the proposed intervention (e.g., because they were a child). Szasz also believed that, even in the case of competent adults who voluntarily seek treatment, psychiatry was primarily a means of controlling people's behavior, bringing the authority of medicine to bear on social judgments of how people chose to live.

Another prominent anti-psychiatrist (though he actually rejected this label) was R.D. Laing (Laing, 1985). Laing emphasized that the aberrant behaviors of people who are in a state of psychosis are actually attempts to communicate with others. This was in opposition to the dominant view, which claimed that a central characteristic of psychotic behavior, including speech, was that it was meaningless. Laing's view was that we should look, not to biology, but to society (especially the family) for an understanding of the causes of mental illness.

Perhaps due to the increasing dominance of the medical model of psychiatry and, especially, of the view that mental disorders are products of alterations in brain function, the anti-psychiatry movement does not have the same public attention and influence it once had. Another possible explanation, and one that is compatible with broadly anti-psychiatry views, is that the anti-psychiatry movement's emphasis on resisting the structures of society was partially a product of its times, reflecting the counter-culture movement of the 1960s and 1970s and partially the beginning of a continued resistance to a solely medicalized approach to mental disorders with the conviction that the social

and cultural systems have a deep impact in the way people experience and respond to disorders of the mind.

Contemporary resistance to psychiatry and psychiatric diagnoses is more likely to come from individuals who have been diagnosed with a mental illness, and had experience with the health care system as a result of their diagnosis, than from clinicians. People who critique psychiatry from this perspective may take a variety of approaches: some are generally accepting of psychiatry but want to ensure that the perspectives of service-users be central in health care practice and policy, and in research. Others are strongly critical of current practices and may call themselves "survivors" of the medical, psychiatric system. Like the anti-psychiatry movement, and sharing some of its motivations and ideas, the psychiatric survivors' movement was also particularly active in the 1970s, and into the 1980s. (It also reflected a broader rejection of medical paternalism and increasing focus on patients' rights.)

Most recently, a number of activists have challenged the claim that psychiatric diagnoses are actually diseases or disorders of any kind. This claim reflects, to some extent, Szasz's view, but whereas Szasz rejected the idea that psychiatric diagnoses should be understood in terms of biology, the current neurodiversity movement embraces the idea that the groups picked out by diagnostic labels are identifiable not just by their experiences and characteristics, but also by shared neural functioning that diverges from typical brain function. The neurodiversity movement says that we should refrain from pathologizing, and perhaps even celebrate, these differences. For example, neurodiversity movement advocates would vehemently oppose your brother or April being diagnosed with ADHD for their behavior and would suggest rather that they vary on attentiveness scale and that we should just embrace their difference as opposed to labeling them with an illness. On the other hand, movements such as *Mad Pride* would encourage April and your brother to own their illness with pride and not be ashamed of it.

The medical model, anti-psychiatry, the survivors' movement, the Mad Pride movement, and advocates of neurodiversity can all be understood as having different philosophical views about the nature of mental disorders and how to think about people who have been diagnosed with one. They differ with regard to their views on the metaphysics of psychiatry—on the nature of mental disorders (or, the term some prefer, mental differences)—and on their ethical and political beliefs about how we should understand and support people who experience these differences. While many of the people who would describe themselves as accepting one of these approaches to understanding psychiatry would be

unlikely to think of themselves as having philosophical views, philosophy can help us to analyze their beliefs and to articulate the similarities and differences among them.

Conclusion

Although there has long been a close link between philosophy and psychiatry, it is only in the past few decades that philosophy of psychiatry has grown significantly as a field of philosophical inquiry in its own right, with its specific set of questions and themes generating interest from both traditional philosophers and mental health professionals, including psychiatrists, psychologists, and social workers. This growth has in part been spurred by changes in psychiatry and psychiatric research. These changes include recent advances in neuroscience, genetics, psychopharmacology, coupled with the publication of the fifth edition of the *DSM (DSM*-5*)*, as well as the National Institute of Mental Health's subsequent initiation of the Research Domain Criteria (RDoC) which offers a new criteria for research by redefining the basic dimensions of functioning (such as fear circuitry or working memory) to be studied across multiple units of analysis, from genes to neural circuits to behaviors. Together, these developments have generated many philosophical and scientific questions. In fact, questions about how to categorize mental disorders form a recurring thread throughout the book—they are central to questions about the definition of mental disorders, the validity and reliability of scientific categories of mental disorders, cross-cultural differences in the properties of mental disorders, responsiveness of the mental disorder criteria to the experience of people with mental disorders, the ethics of psychiatric diagnoses and treatment, and the nature of psychiatric research as a scientific enterprise. We believe that all of these questions can be best answered through collaborative work by philosophers, scientists, and medical practitioners, generating further interest in philosophy of psychiatry as a discipline in its own right.

We envision this book as being useful to a broad audience. It can be used as an introductory textbook for students new to both philosophy and psychiatry, or to provide an overview of philosophy of psychiatry for students, teachers, researchers, or practitioners in one of the two areas who want to learn more about the other. Some of the chapters serve as introductory overviews to an area of philosophy of psychiatry; in addition to this chapter and the following chapter by Mona Gupta, these include the chapters by Jesse Butler and by Ginger

Hoffman (both on philosophy of mind), by Claire Pouncey and Jon Merz (on informed consent), and, in the part on metaphysics, the chapters by Christian Perring (on autonomy and free will) and by Peter Zachar (on psychiatric classification). At the same time, each chapter in the book is an original piece of scholarship, written not only to introduce, but to advance the conversation on its topic. Taken as a whole, the book is intended to contribute to philosophy of psychiatry, as a vibrant area of philosophical inquiry, by showing both that psychiatry raises important philosophical questions and that it can shed light on perennial philosophical problems.

Because our focus has been contemporary philosophy of psychiatry, we had to leave out important moments in the history of psychiatry as well as the development of philosophy of psychiatry, the materials on which, we hope, will be used as supplements to this book. To mention a few, this history includes the emergence of psychiatry as a medical specialty in 1800s with a specific concern over individuals who are confined to asylums or hospitals; the contributions of the neurologist Sigmund Freud to the understanding of mental disorders through his theory of psychoanalysis, according to which mental disorders emerge due to interaction of conscious and unconscious elements in the mind and can be successfully treated by bringing repressed fears and conflicts into the conscious mind through techniques such as dream interpretation and free association; the subsequent concerns over psychoanalysis not being a scientifically rigorous paradigm to investigate mental disorders due to its theoretical presuppositions about the human mind, for example, the unresolved sexual tensions of childhood causing adult psychosis; and the rise of scientific psychiatry starting with 1970s. The reader may find some resources in the annotated bibliography to pursue these topics further.

Our second reflective exclusion in the book is the first person perspectives of those suffering from mental disorders. Thanks to the increased awareness about the effect of mental disorders on people's lives, attempts to reduce stigma, and increased ease of publishing through the advances in internet publishing, many individuals with mental disorders recount their experiences with mental disorders. We believe that engagement with these perspectives is necessary for those who are interested in philosophy of psychiatry. The first person accounts offer a unique window for understanding the person with a mental disorder, and they illustrate the commonalities and the differences among the experiences of individuals coming from diverse backgrounds as substantially illustrated in Capponi (1992), Cheney (2009), Jamison (1995), Reaume (2009), Saks (2007), White (2008), and Romme and Escher (1989, 1993). These accounts

afford a deeper appreciation of how psychopathology impoverishes a person's relationship with herself and her community. We left these forms of inquiry out of the scope of this book for reasons of space. However, we expect and encourage that such readings supplement the content of this book. In addition important philosophical work on the importance of first person perspectives in mental health research and treatment may also be used to explore these topics further (e.g., Ankeny, 2017; Flanagan, 2013; Kleinman, 2012; Radden and Varga, 2013; Tekin, 2011, 2014). The reader may find more about these resources in the annotated bibliography to pursue these perspectives further.

What you find next is Mona Gupta's chapter "What Is Psychiatry?," which addresses the question of what psychiatry is, distinguishing it from other areas of medical practice, other health care professions, and from psychology. The remainder of the book is divided into parts corresponding to traditional areas of philosophy: philosophy of mind, phenomenology, philosophy of science, ethics, social and political philosophy, metaphysics and epistemology, and philosophy of medicine.

The part entitled "Philosophy of Mind and Psychiatry" focuses on how the philosophical debate on the mind-body problem illuminates central questions in psychiatry, through examples of mental disorders such as autism and psychopathy. It starts with Jesse Butler's chapter, "The Type/Token Distinction and the Understanding of Mental Phenomena in the Biopsychosocial Model," which maps the biopsychosocial model in psychiatry on to the traditional questions about token and type identity theory in philosophy of mind. It is followed by Ginger Hoffman's chapter, "Frequently Asked Questions about Mental Disorders," which focuses on the implications of the relationship between the mind and the brain for the understanding of mental disorder. Next, Erick Jose Ramirez evaluates the role of emotions psychopathy and autism in "Psychopathy, Autism, and Basic Moral Emotions." Finally, Lana Kühle draws on recent discussions of embodiment in philosophy of mind to discuss anorexia nervosa in "Anorexia Nervosa and the Embodied Mind."

The part "Phenomenology and Psychiatry" starts with Anthony Vincent Fernandez's chapter, "Merleau Ponty and the Foundations of Psychopathology," which shows the influence of Merleau-Ponty's work on understanding the first person experience of mental disorder. The following chapter by Constantin Mehmel, "Transformation through Dialogue: Gadamer and the Phenomenology of Impaired Intersubjectivity in Depression," provides an alternative phenomenological approach to understand the nature of intersubjectivity in depression.

The "Philosophy of Science and Psychiatry" part starts with Jonathan Y. Tsou's "Philosophy of Science, Psychiatric Classification and the *DSM*." Tsou focuses on various philosophical questions in the development of the most influential scientific taxonomy of mental disorders, namely, the *DSM*. Aaron Kostko's chapter "Inductive Risks and Psychiatric Classification" evaluates the role epistemic and non-epistemic of values in scientific psychiatry. In "Causal Explanation in Psychiatry" Tuomas Pernu elaborates on various approaches to causation in the scientific research on mental disorders and shows the practical implications of these different approaches. This part ends with Ami Harbin's application of some of these themes to research on trauma in her chapter "Trauma-Informed Psychiatric Research."

Claire Pouncey and Jon F. Merz open the "Ethics and Psychiatry" part with "Informed Consent and Psychiatric Treatment and Research" where they highlight the importance and complexity of informed consent in various psychiatric settings. Next, Kelso Cratsley elaborates on coercion in psychiatry in his chapter "The Ethics of Coercion and Other Forms of Influence." In "Voice, Silence, and Listening Well" Nancy Nyquist Potter elaborates on the ways that people in mental distress experience themselves as silenced or otherwise not heard. Michelle Ciurria's chapter "Responsibility as a Dialogical, Care-Based Interaction" develops an account of responsibility that is responsive to the complexity of the experiences of people with mental disorders. Matthew Ruble concludes this part with "Philosophers, Psychopaths, and Neuroethics" where he furthers the analysis of responsibility in mental disorder context by engaging with the question of whether it is morally permissible to blame someone for contracting an illness that renders the individual a morally blind agent.

The part "Social and Political Philosophy and Psychiatry" starts with Robert Chapman's discussion of neurodiversity in his "Neurodiversity and Its Discontents," which provides a non-medical perspective on understanding mental disorders by construing them instead as mental differences. It is followed by Devonya Havis and Melissa Mosko's "Managing Individuals and Populations through Psychiatric Classification" in which they apply Michel Foucault's concept of biopower to the state-enforced sterilization of poor women and women of color, as well as state-sanctioned lobotomies of women in the twentieth-century framework. Bryan Mukandi in "The 'North African Syndrome'" draws on literary texts on psychopathology to examine Frantz Fanon's 1952 essay, "The 'North African Syndrome.'"

Christian Perring opens the "Metaphysics, Epistemology and Psychiatry" part with his chapter "Mental Disorder, Free Will, and Personal Autonomy" where he

discusses the influence of mental disorders on individual's free will and personal autonomy. It is followed by Peter Zachar's "Metaphysical Problems in Psychiatric Classification and Nosology" where he relates the core questions in psychiatry to the traditional questions in metaphysics. Sam Fellowes, in a similar fashion, connects the core epistemological questions about realism and antirealism in philosophy to psychiatry, in his chapter "Realism, Antirealism and Psychiatric Classification." Anke Bueter concludes this part with "Social Epistemology and Psychiatry" with a discussion of the notion of testimony and its import and application in psychiatry.

The final part, "Philosophy of Medicine and Psychiatry," contains two chapters that connect psychiatry to the field of medicine at large. Hane Htut Maung compares the role of diagnosis in medicine and psychiatry in his chapter "The Functions of Diagnoses in Medicine and Psychiatry." Reinier Schuur concludes this part with a discussion of the importance of conceptual analysis in philosophy of psychiatry in his chapter "Mental Health and Illness: Past Debates and Future Directions."

Overall, the book provides an overview of work in contemporary philosophy of psychiatry, with each chapter providing an "opinionated introduction" to the relevant issues, and we hope that it inspires further interest in the important and urgent topic of mental disorder.

References

Ankeny, R. (2017), The Role of Patient Perspectives in Clinical Case Reporting, pp. 97–112. In *Knowing and Acting in Medicine*. R. Bluhm, ed. New York: Rowman and Littlefield.

Capponi, P. (1992), *Upstairs in the Crazy House: The Life of a Psychiatric Survivor*, Toronto: Viking.

Cheney, T. (2009), *Manic: A Memoir*, New York: HarperCollins Publishers.

Flanagan, O. (2013), Identity and Addiction: What Alcoholic Memoirs Teach, pp. 865–888. In *The Oxford Handbook of Philosophy and Psychiatry*. K. W. M. Fulford, Martin Davies, G. T. Richard George Graham Gipps, John Z. Sadler, Giovanni Stanghellini, and Tim Thornton, eds. Oxford: Oxford University Press.

Jamison, K. (1995), *An Unquiet Mind: Memoirs of Moods and Madness*, New York: Knopf.

Kleinman, A. (2012), The Art of Medicine: Culture, Bereavement, Psychiatry, *The Lancet*, 379(9816):608–609.

Laing, R. D. (1985), *Wisdom, Madness and Folly: The Making of a Psychiatrist 1927–1957*, London: Macmillan.

Radden, J. and Varga, S. (2013), The Epistemological Value of Depression Memoirs: A Meta-Analysis, pp. 99–116. In *The Oxford Handbook of Philosophy and Psychiatry*. K. W. M. Fulford, Martin Davies, G. T. Richard, George Graham Gipps, John Z. Sadler, Giovanni Stanghellini, and Tim Thornton eds. Oxford: Oxford University Press.

Reaume, G. (2009), *Remembrance of Patients Past: Patient Life at the Toronto Hospital for the Insane, 1870–1940*, Toronto: University of Toronto Press.

Romme, M. and Escher, S. (1989), Hearing Voices, *Schizophrenia Bulletin*, 15(2):209–216.

Romme, M. and Escher, S. (1993), *Accepting Voices*, London: MIND.

Russell, B. (1912), *The Problems of Philosophy*, Oxford: Oxford University Press.

Saks, E. (2007), *The Centre Cannot Hold: My Journey through Madness*, New York: Hyperion.

Szasz, T. S. (1961), *The Myth of Mental Illness: Foundations of a Theory of Personal Conduct*, New York: Harper & Row.

Tekin, Ş. (2011), Self-Concepts through the Diagnostic Looking Glass: Narratives and Mental Disorder, *Philosophical Psychology*, 24(3):357–380.

Tekin, Ş. (2014), Self-Insight in the Time of Mood Disorders: After the Diagnosis, Beyond the Treatment, *Philosophy, Psychiatry, and Psychology*, 21(2):139–155.

White, R. (2008), Finding My Mind, pp. 44–69. In *Breaking The Silence: Mental Health Professionals Disclose Their Personal and Family Experiences of Mental Illness*. S. Hihnshaw, ed. Oxford: Oxford University Press.

2

What Is Psychiatry?

Mona Gupta

The basic problem

Psychiatry is the branch of medicine concerned with the diagnosis and treatment of mental disorders. Underlying this seemingly simple definition are several questions. To what extent does the practice of psychiatry fit within the medical paradigm: from the medical model of disease, to medicine's institutional location, to its daily tasks of diagnosis and treatment? Do the various controversies that have surrounded the field suggest that it is different in some fundamental way from other medical specialties, such as cardiology or general surgery? And how does psychiatry differentiate itself from non-medical fields concerned with mental disorders, like psychology and social work?

Typically, medicine is understood to be about disease and, more specifically, the diagnosis and treatment of disease. Disease refers to a pathological structure or function (or both) in the body. Diagnosis refers to the correct identification of the pathology that generates a person's symptoms, while treatment refers to the use of various interventions to cure disease by correcting the pathology (e.g., antibiotics which kill the bacteria causing pneumonia) or to palliate the symptoms caused by disease (e.g., anti-inflammatory drugs which reduce the pain and swelling caused by osteoarthritis without curing the disease process which causes pain and swelling).

While there is extensive debate about the precise meaning of the phrase "medical model" of disease (Murphy, 2011: 426–427) we can pick out some common features. First, disease is ultimately realized in disturbances in bodily processes; second, a correct diagnosis is the best starting point for successful treatment and accurate prognostication; third, correction of the disturbed bodily process will likely lead to improvement in the symptoms of the disease; and fourth and finally, biological sciences and epidemiology are

needed to understand disease rather than social sciences, humanities, or other disciplines.

Our current understanding of mental disorders diverges from this characterization. First, they cannot be defined in reference to an underlying pathology given that for the vast majority of the entities considered to be mental disorders; there is no specific, identifiable pathophysiologic process. Instead mental disorders are defined by categories which represent groups of symptoms that tend to cluster together. The symptoms are articulated through operationalized criteria for each disorder. In North America, the classification scheme of all the disorders is described in the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders*, currently in its fifth revision *(DSM-5)* (APA, 2013). And because similar symptoms may cluster with different disorders, a single, correct diagnosis is less common than overlapping categories or comorbidity. For example, the difficulty sleeping and concentrating that may be part of a major depressive disorder can also be symptoms of a generalized anxiety disorder. If a patient experiences both depressed and anxious mood, it can be difficult to know whether a person has a depressive disorder, an anxiety disorder, or both at the same time. Officially, a person whose situation meets the criteria for a category should be diagnosed with that condition. From the person's experiential point of view, her state of being may feel like one phenomenon, but she may end up with several diagnoses. Second, psychiatric treatments are not oriented toward the correction or palliation of a disturbed bodily process but toward alleviation or elimination of the symptom groups (psychotic symptoms, depressive symptoms) that make up the diagnostic criteria. Diagnostic accuracy, that is, identifying the precise *DSM* category which applies to a given patient, is therefore less important than in other areas of medicine because diagnosis does not necessarily determine treatment. Finally, while much has been made of the importance of exposing physicians in training to the humanities and social sciences in order to improve their capacity to understand their patients' illness experiences, these disciplines do not further the understanding of disease per se. Type 1 diabetes mellitus is an absolute deficiency of insulin regardless of how the sufferer experiences it, her cultural background, or her personal beliefs. Illness experience varies but pathophysiology defines physical disease more or less completely. Contrarily in psychiatry, the illness experience *is* the disease; therefore, the humanities and social sciences may be necessary to understand the disorders themselves in light of the variation that exists in idioms of distress across time and geographical space. For example, a mental state or experience that might fulfill criteria for major depression in North America might be

understood differently in another country or culture. Disciplines that explore local meaning are necessary in order to understand what the expression of a given experience or mental state means; that is, is it a medical condition, is it a (normal) distress signal in a given circumstance, etc. In other words, the non-biological sciences are also necessary to understand disease.

In part, it is in recognition of these distinct aspects of psychiatric disorders that psychiatrists have aligned themselves with the biopsychosocial model rather than the biomedical model. First advanced by the internist George Engel in the 1970s, the biopsychosocial model posits that all diseases are the result of complex interactions occurring biological, psychological, and social levels (Engel, 1977). These interactions cannot be reduced to one basic level nor can one understand the problem fully by focusing only on one level. Indeed, Engel explicitly contrasted this model with the standard biomedical model. Although it has its critics (Ghaemi, 2009), the biopsychosocial model is still considered *the* major explanatory model in psychiatry (ABPN, 2011; RCPSC, 2015). That is, psychiatrists are expected to explain their patients' problems in reference to the biopsychosocial model rather than the medical model. Yet, adherence to a biopsychosocial approach seems to put psychiatry at odds with its parent discipline of medicine.

Who are psychiatrists?

Psychiatry's practitioners, psychiatrists, are medical doctors. This means that all psychiatrists first complete medical school (which is a university program at either the undergraduate or graduate level, depending on the country). They then pursue an additional multi-year period of training (residency) to become specialists in psychiatry. The duration of residency training in psychiatry lasts several years, which is similar to residency training in other areas of medicine. During specialty training, residents are exposed to different types of psychiatric practice by clinical setting (hospital vs. outpatient clinic), by population (children, the elderly), and by disorder type (chronic psychotic disorders).

The fact of being physicians is the most obvious distinction between psychiatrists and other clinicians who work in the domain of mental health, such as psychologists, nurses, social workers, occupational therapists, among others. It means that psychiatrists can perform certain clinical acts and have certain responsibilities specific to physicians that other professional groups do not have and cannot assume. The first and most obvious of these is the prescription of medication.

Other physicians, such as family physicians, can also prescribe psychiatric medications, but non-physicians may not.[1] Another aspect of practice which is specific to psychiatry includes the consideration of legal issues such as involuntary hospitalization and/or treatment for mental disorders[2] or assessment of criminal responsibility of a mentally ill person. Finally, there are tasks that psychiatrists may undertake which include providing a psychiatric diagnosis, prognosis, and treatment plan for various administrative and/or legal documents such as letters in support of applications for financial benefits or other social programs.

Psychiatric work is usually conducted within the institutional setting of medicine which includes hospitals and medical clinics of various types. In recent years, a variety of community treatment models have been developed to make receiving care less burdensome for some patients and to offer clinicians the opportunity to assess their patients in the real contexts in which they live. This might include mobile crisis units that make home visits, or assertive community treatment teams based in shelters where they can meet the patients on-site. Despite the fact that these types of services are offered outside institutional boundaries, the psychiatrist's role retains its medical and medicolegal character in that the psychiatrist's task is symptom management with medications and the evaluation of safety to self and others. Usually such work is conducted in an interdisciplinary manner involving clinicians from other disciplines such as nursing and social work and it is often their role to ensure that patients' homes are safe and hygienic and that patients engage in some community activity.

Regardless of the practice setting, psychiatrists, like other medical specialists, increasingly work together in teams of other professionals who bring their disciplinary perspectives and skill sets to the understanding of mental illness. One such perspective is that of the clinical psychologist. Clinical psychologists are not physicians, but instead obtain graduate-level degrees (masters or PhDs) and receive clinical training within these programs. Certain psychologists undertake masters or PhD programs with no clinical training in order to pursue research and/or teaching. Some clinical psychologists assess patients and provide psychotherapy (to be discussed in more detail in a later section) while others focus on the administration and interpretation of formal psychological tests.

Apart from the distinctions in professional training and permitted clinical acts, do psychiatrists and psychologists view mental disorders differently? The majority of psychiatrists and psychologists would agree that what we call the mind is a result of the functioning of the brain. That is, a functioning brain is necessary for both normal and abnormal mental processes. And both psychiatrists and psychologists increasingly adhere to the view that the final pathway of mental disorders has to be preceded by some kind of dysfunction within individual

brains—even if there are more distal psychological or social causes (Schwartz et al., 2016). Psychologists have developed various non-biological theories of disorder (such as cognitive theories of depression, according to which a condition like major depression results from the evolution and activation of a web of distorted cognitions—thoughts, beliefs, and attitudes—which lead to consistently pessimistic and erroneous interpretations about the self, the world, and the future) but these do not preclude biological reduction. Instead, their proponents choose to emphasize the explanatory resources of the cognitive model while remaining open to the idea that they are underscored by biological changes. The medical model objectifies disease and renders the doctor-patient relationship interchangeable. That is, in the medical model, a disease is roughly independent of the person who has it and a treatment roughly independent of the person who delivers it. The development of standardized diagnostic and symptom assessment scales as well as manualized therapies represents this point of view and is being led by research efforts in psychology. Psychiatrists do not reject such developments but the driving force for empirically validated and supported evaluation and treatment methods comes from psychologists. Interestingly, in these respects, psychology seems to adhere more closely to the medical model than does psychiatry.

Several other professionals can be involved in mental health teams and programs, such as social workers, occupational therapists, and nurses. Social workers receive their training at the university level and master's degrees are often required for clinical positions. Occupational therapy training has transitioned from the undergraduate to the graduate level, meaning that licensed clinical practice in this domain requires a graduate-level degree in occupational therapy (undergraduate degrees still exist but alone are insufficient for clinical practice). The field of nursing is in a process of transition from non-university level training to university level training and higher level positions require Masters degrees. Each professional plays a specific role in care. For example, social workers are familiar with existing social and community resources and frequently assist their patients in accessing financial resources and adequate housing. Occupational therapists may help patients to re-engage in activities that have been abandoned or undeveloped due to illness. Nurses often provide case management and care in multiple domains such as assessing the impact of medication, or implementing changes from psychotherapy such as developing alternative interpretations of situations or reducing avoidance of social interaction. In the inpatient setting, the team typically works together with each person providing input as it concerns their own areas of expertise and depending on the patient's needs. In the outpatient environment, patients often meet with a primary caregiver who

is leading the care plan at a given point in time, and whose discipline varies depending on the program or practice setting, while the others provide input as needed. All mental health professionals, including psychiatrists, may or may not be formally trained and competent to practice psychotherapy.

Diagnosis and treatment

Diagnosis

What do psychiatrists do in their daily practice? While practice settings differ, there are certain tasks which characterize the clinical encounter. These include principally diagnosis, and treatment planning and oversight (including both prescribing of medication and psychotherapy). When a patient is new to a psychiatrist, a certain amount of time is spent on the diagnostic evaluation, namely, the process by which psychiatrists categorize the patient's problem(s). Ordinarily this is done in accordance with a formal diagnostic classification scheme. Making a psychiatric diagnosis usually involves questioning the patient about the symptoms that make up the operationalized criteria for a given category. The answers to these questions may be supported by the scores and answers obtained in symptom rating scales. Psychiatrists also conduct a mental status exam which is meant to provide an objective rendering of various domains of the patient's mental functioning such as his perception, thinking, judgment, cognitive function. These will be evaluated in more or less detail depending on the patient's problem. For example, it may be more important to explore a patient's cognition if she is experiencing memory loss than it would be for someone who is having panic attacks and has no obvious difficulties in cognitive function. Inevitably, all of this information must be triangulated with the psychiatrist's experience of the patient. That is, the psychiatrist also learns about the patient as a person intersubjectively; thus, the diagnostic evaluation contains subjective, objective, and intersubjective elements which enable the psychiatrist to offer up a psychiatric diagnosis.

The information solicited during a psychiatric evaluation is usually organized following a specific structure and documented as such. This adheres to a structure similar to what occurs in general medical evaluations with some additional or expanded categories of information.

Table 1 identifies the categories of information and the types of information gathered in each category. The goal of this part of the psychiatric evaluation is

Table 1 Structure of a General Psychiatric Evaluation

Category	Description
Identification	Age, housing status, marital status, parental status, employment, source of revenue.
Sources	Identifies the sources of information for the evaluation which could include the patient, family members or friends, other clinical staff, medical chart, referral letter or consultation request.
Chief complaint or reason for referral (CC or RFR)	This includes the primary reason the evaluation is being sought and can be documented from the patient's point of view and the referring source's point of view when there is a specific referral being made.
History of present illness (HPI)	A chronological summary of the development of symptoms or problems included in the CC or RFR. The purpose of the HPI is to organize the summary such that it points toward a short list of possible diagnoses.
Functional enquiry	A survey of other common psychiatric symptoms associated with the symptoms included in the CC or RFR.
Collateral history	The events leading to the evaluation as told by a third party, such as a family member or another clinician who knows the patient.
Past psychiatric history	Details of a person's past encounters with mental health professionals, psychiatric hospitalizations, psychiatric medication use, suicide attempts.
Substance use	Consumption of legal and illegal non-pharmacological substances, including caffeinated products.
Forensic history	Details regarding the person's previous encounters with the criminal justice system.
Medical history	Conditions and treatments in the past, present, or that are currently under investigation, including sexual history.
Medications	Include prescribed, over the counter, herbal, and homeopathic remedies.
Allergies	Documented allergies as well as intolerances.
Family psychiatric history	Details of any family members' past encounters with mental health professionals, psychiatric hospitalizations, psychiatric medication use, suicide attempts.
Family history	Genogram of the family identifying the index patient (the person being evaluated).
Personal history	Details of the person's life trajectory from birth to the present. Includes family life, personal relationships, schooling, work, spirituality, interests, and important life events.

Sadock and Sadock (2011: 227–232)

to fully characterize the current problem, the medical, psychiatric, and personal context in which it is situated and to point toward potential diagnoses. The structure outlined below reflects a general psychiatric evaluation. Additional aspects may be included in specific sub-disciplines of psychiatry. For example, an evaluation in child psychiatry may typically include more information about the child's parents than one would find in an adult patient's report.

In addition to gathering this information, psychiatrists also undertake a mental status examination (MSE) during the evaluation. The MSE is intended to be an objective portrait of those domains of mental life considered to be those most relevant to psychiatric disorders. The psychiatrist assesses these domains in real time. For example, if the patient is calm during the evaluation, this is what is documented in the MSE even if he was agitated yesterday. His report of being agitated would be included in the HPI. Table 2 summarizes the principal domains evaluated in the MSE.

Table 2 Domains of the MSE in Psychiatry

Domain	Description
Appearance	Includes observations of posture, grooming, hygiene, physical illness.
Behavior	Observations of psychomotor activity including such items as tics, mannerisms, gait disturbance, agitation, or retardation.
Eye contact	Presence and quality of the person's gaze.
Attitude	This item refers to the patient's attitude toward the psychiatrist and might include such qualities as cooperative, contemptuous, apathetic, evasive, guarded, hostile.
Speech	Refers to the rhythm, rate, and tone of the person's spoken language.
Mood	Refers to the person's description of her emotions.
Affect	Refers to the person's expression of her emotions and may or may not be congruent with her stated mood. Other aspects of affect include its range, its amplitude, and its reactivity.
Perception	Refers to the presence or absence of hallucinations or illusions.
Thought content	Refers to what the person is thinking about during the evaluation. This includes the themes raised by the person as well as an assessment of the presence of absence of obsessions, overvalued ideas or delusions (which include those that paranoid, grandiose, guilty, nihilistic, somatic, or erotomanic), and violent ideation toward the self or others.
Thought process	The process of thought refers to the manner in which the person expresses his ideas and associations between ideas.

Table 2 (*Continued*)

Domain	Description
Cognition	Includes alertness, orientation (person, place, and time) concentration, memory, (immediate, recent, and long term) calculation, fund of knowledge, abstract reasoning. All of these must be considered in light of the person's highest-level education.
Insight	Insight is the extent to which the person is aware of and understands her problems.
Judgment	Judgment refers to a person's ability to act appropriately to the circumstances she is in. This includes the ability to exercise self-restraint and predict outcomes of behavior.

Sadock and Sadock (2011: 232–237)

Clinical case formulation

What is the purpose of a clinical encounter in medicine? Patients typically attend a physician with a problem causing them some type of discomfort, disability, or distress. They are usually in search of two things: relief from their suffering (treatment) and an explanation for the problem (diagnosis). Because current psychiatric diagnoses represent groupings of symptoms, rather than an underlying pathophysiological process, they cannot offer the usual type of explanation that diagnosis conveys. Instead, psychiatrists must be able to synthesize the information and observations they have gathered about the person in order to generate an explanation of why the person is experiencing the problems she is experiencing. This is known as the clinical case formulation. The formulation is considered to be a necessary component of diagnosis and treatment planning.

A formulation attempts to explain something about a person, not about a diagnosis or a disease. Indeed, the authors of *DSM*-5 note that the purpose of a *DSM* diagnosis is to serve the elaboration of a clinical formulation (APA, 2013: 19). That is, the diagnosis helps to understand the person but does not represent the person, nor their problem. According to Sturney, formulations can be told from individual theoretical perspectives, from the integration of two or more of these, or by borrowing from diverse sources (eclectic approach) (Sturney, 2009: 8). Regardless of theoretical approach, formulations have common elements. First, they abstract out key features of the case; second, they tie together the onset, development, and maintenance of the problem(s) and should link these ideas to the treatment that should grow out of and relate to the formulation. Third, they are tentative and subject to modification in light of new information;

and fourth, they should predict individually designed treatments that will be more effective than treatments that would otherwise have been implemented (Sturney, 2009: 8).

The biopsychosocial model of mental disorder permits the psychiatrist to offer an explanation that integrates biological, psychological, and social causal factors as appropriate. Biological factors may include a family history of a condition, psychological factors may draw upon psychological theories to explain specific aspects of the problems (defensive patterns, cognitive distortions, etc.) while social factors may point to the difficult circumstances (poverty, abuse) which may be contributing to the person's problem. Although the case formulation seems to deviate from the medical model of disease by explicitly permitting non-biological factors to explain the "disease," in fact, it shares the overall aims of medical diagnosis, namely, to offer a causal account of the person's problem that will both explain its development and point to avenues for treatment.

Treatment

Once a diagnosis and formulation have been established, psychiatrists help to plan treatment whether they deliver that treatment themselves, or they participate in the evaluation of its effectiveness for the patient. A treatment plan may involve the psychiatrist as the central provider or the psychiatrist offering part of the treatment in parallel with another provider (e.g., the psychiatrist prescribes medication and assesses its impact while a psychotherapist provides psychotherapy). In the latter case, the psychiatrist's role usually involves seeing the patient periodically, the frequency depending on the severity of the problem and the patient's participation in parallel resources, enquiring about symptoms as they evolve (or not) throughout the treatment process, and reevaluating the patient's mental status. Or psychiatric treatment may involve a psychiatrist acting as a consultant who sees the patient infrequently and provides treatment recommendations to other providers who deliver and oversee their use. These other providers may work together with the psychiatrist in a team or may be professionals who accept referrals from psychiatrists or self-referrals from patients.

Intervention in the treatment of mental disorder can take one of two broad forms: biological and psychological. The use of these terms does not necessarily refer to the nature of the interventions, nor mechanisms of action (biological treatments can have important psychological dimensions and psychological treatment may influence biological processes) but is a convention often found in psychiatry textbooks and the scientific literature thus, it will be maintained here.

Psychiatrists may use one or both such types of interventions in their clinical practice.

All psychiatrists are trained to understand the indications, side effects, and potential drug interactions of psychopharmacological agents. There are at present, many medications available on the market to treat psychiatric disorders. Clinically, these agents tend to be grouped together according to the symptoms they target, and then further subdivided by biochemical structure or mechanism of action. There are, for example, antidepressants, anxiolytics, antipsychotics, and mood stabilizers. Within the clinical grouping of the antidepressants, there are several subdivisions: the tricylics (so named because of their chemical structure), the selective serotonin reuptake inhibitors (so named because of a specific action at the neuronal synapse), and several others. The antipsychotics are subdivided into the so-called typical (first generation) and atypical (second generation) antipsychotics where typicality refers to certain mechanisms of action observed in animal experiments. The mood stabilizers include agents of diverse structures, most of which are also antiepilepsy drugs with the exception of lithium, a naturally occurring salt. Finally, the anxiolytics include the large group of biochemically related benzodiazepines, several of the antidepressants which are used to treat anxiety, certain atypical antipsychotics, and a small number of other agents. All of these agents have their own pharmacologic profile, including side effects and drug interactions. Similar to other physicians, psychiatrists' work usually involves prescribing medications.

Among those clinicians who work in mental health care, psychiatrists are also uniquely qualified to prescribed and administer brain stimulation or neuromodulation techniques, including electroconvulsive therapy (ECT), transcranial magnetic stimulation (TMS), vagal nerve stimulation, and deep brain stimulation, although these latter three techniques remain experimental and are primarily available in research settings under specific circumstances. The therapeutic effect of ECT requires the induction of a bilateral generalized seizure during which the patient is anesthetized and paralyzed (therefore they do not experience the actual physical movements of someone having a generalized seizure, instead the seizure happens at the level of brain activity only). ECT is typically prescribed for a set number of treatments (i.e., 6 or 12). Less commonly it is given as a maintenance treatment; that is, the person receives it on a regular basis (e.g., weekly) in order to prevent relapse of a mental disorder. The primary clinical indication for ECT is severe major depression (psychotic or not) and less commonly severe mania, acute, severe symptoms of schizophrenia, and catatonia. Providing ECT is more resource intensive

than prescribing medication as it involves hospital resources and personnel, including an anesthetist, anesthesia medications, and equipment, as well as a staffed recovery room with monitoring equipment. TMS involves passing a magnetic field over the scalp and cranium which depolarizes superficial cortical neurons to stimulate such neurons thereby relieving symptoms of depression. Vagal nerve stimulation involves repeated delivery of electric pulses to the left vagus nerve by a pulse generator implanted in the left chest wall. Finally, deep brain stimulation requires neurosurgical intervention in that it involves passing a fine wire into specific brain regions through a hole in the skull. Stimuli can be transmitted to the wire by a pulse generator implanted in the chest wall and connected subdermally (Sadock and Sadock, 2011: 1117–1126).

A subset of psychiatrists also develop competence in the use of non-pharmaceutical therapies, primarily psychotherapies of various types for individual, couples, families, or groups. Whether or not they obtain such skills depends on the training requirements of their national certification organizations—some require practice-level competence in psychotherapy, others do not. Some practitioners also choose to develop such skills after residency training. While there are dozens of specific therapies, there are a small number of broad traditions in which most psychotherapies fall. The psychodynamic tradition is a descendent of psychoanalysis and shares some of its theoretical commitments, including the existence of the unconscious, the necessity of understanding childhood experiences in order to explain present problems, and the central role played by intrapsychic conflict and defense mechanisms in understanding current symptoms. The dynamic therapies are liberatory: if people can develop conscious understanding of themselves, they can then choose how they want to live rather than be trapped by unconscious impulses, conflicts, and patterns.

Cognitive-behavior therapy by contrast maintains a specific focus on the present, on symptom reduction, and on the conscious thoughts that lead to the development of pathological feelings and behaviors. It is a technology of change. Finally, the systemic therapies all involve a commitment to viewing the problem as residing within a group of people (e.g., a family) rather than an individual and in which every member plays a role. In other words, problems—and potential solutions—have to be understood from the point of view of the whole system—how it creates them and how it will respond (Sadock and Sadock, 2011: 940–943) Apart from some family physicians who may be trained in psychotherapy no other physician group routinely *delivers* non-pharmacological or non-surgical therapeutic interventions. Some specialties work closely with non-

medical clinicians who do deliver such treatments. For example, physiatrists will prescribe physiotherapy but do not themselves deliver it.

Diagnosis and treatment tasks vary somewhat in specialized areas of psychiatry. For example, child psychiatrists will almost certainly spend a portion of their appointment time meeting with the child's (the patient's) primary caregivers. Forensic psychiatrists carry out similar tasks, but with an eye toward the questions asked of them by the legal system (e.g., whether the person is fit to stand trial, or criminally responsible for an act committed). Some psychiatrists will also have to engage in medico-legal tasks such as attending review board hearings for patients who challenge their involuntary hospitalizations or treatment or administrative panels whose role is to determine whether someone can continue live independently or requires supervised housing.

Challenges facing psychiatric practice

Notwithstanding the principal problem faced by all physicians—that there are many patients whose problems are resistant to available therapeutic interventions—there are certain challenges that are specific to psychiatric practice. In this section I will discuss two that arise directly from the issues discussed in this chapter: (1) the clinical consequences of the uncertain nature of psychiatric diagnoses and (2) the impact of the medical model on psychiatric work.

The clinical consequences of the uncertain nature of psychiatric diagnoses

As discussed above, the purpose of diagnosis is to identify the pathophysiological process that is causing the patient's symptoms, so that this process can be acted upon by treatments to cure the disease or palliate the symptoms. However, diagnosis also serves other purposes. It enables employers, courts, insurers, and others to distinguish sickness from non-sickness. Typically, a sick or disabled person is entitled to certain types of advantages or accommodations. For a worker this may include paid disability leave; for a student it may include extensions for submitted work or special conditions during formal evaluations. The physician is called upon to determine whether (a) the pathological process corresponds to the person's stated illness or disability and (b) whether the underlying pathology does give rise to certain limitations. Because psychiatric

conditions do not refer to underlying pathology, the psychiatrist cannot make the same kind of determination.

Because the *DSM* classification scheme includes social or occupational dysfunction as an essential component of the definition of most diagnoses, as well as allowing catch-all categories for clinical situations in which a person does not meet the full criteria of a syndrome, a great many situations can technically fulfill the criteria of a mental disorder. The uncertainty about whether something is or is not a mental disorder leads to a certain degree of elasticity in diagnosis in which categories can expand to include a wide range of cases and new categories can flourish. This makes it very difficult for psychiatrists to exclude persons as cases of mental disorder when distress and dysfunction are great. But surely not every case of distress or dysfunction ought to be considered a mental disorder. Or should it?

A medical diagnosis such as asthma can help to illustrate the issue at stake. If a worker were to claim that he could not fulfill a certain task, such as work in a certain environment because it worsened his asthma, a physician would not only have to make the diagnosis of asthma, that is, to establish that asthma was present using measures apart from the patient's reports of symptoms, but also have to determine if the nature and severity of the person's particular disease prevented him from doing the task. Nature and severity would be determined by comparing the patient's disease to the population of patients with the same disease by means of comparing various measurements of their disease states. For asthma, this may include measures such as forced expiratory volume, vulnerability to the presence of a known trigger, frequency of exacerbations, numbers and doses of medications required to reduce the exacerbation. The doctor has multiple sources of information to inform a clinical judgment of sick or disabled, including sources that go beyond the person's self-assessment.

Take by contrast, a patient whose workplace is undergoing major organizational changes: colleagues are being laid off, departments are being closed, and her job tasks are shifting toward ones she finds uninteresting. The managers responsible are unsympathetic toward affected employees and wait until the last minute to share information with them. At work, the patient feels sad and is less attentive to her work. She does not want to socialize with colleagues and does not even feel hungry enough to eat her lunch. Sometimes she sleeps poorly if she is upset about what happened at work the day before. Outside of work, she feels better but she does worry about what is going to happen at her workplace. She avoids seeing her friends and family as she does not want to talk about what is going on at her job. This patient is very distressed, she says she cannot work. In this

case, the physician has to make a determination of sick or disabled with respect to her job, without the benefit of information beyond the patient's own self-assessment. The issue here is not that the patient cannot be trusted but rather the physician is being asked to make a clinical judgment when it is not at all clear that such a judgment requires clinical abilities. Is this patient suffering from a mental disorder? The uncertain nature of psychiatric diagnoses would allow the psychiatrist to argue either side. The determination of sick or disabled seems to depend heavily on the psychiatrist's values about who should be accorded leave from work and under what circumstances.

There are diagnoses in other areas of medicine where the same type of problem arises, particularly for conditions in which pain is the central debilitating symptom (e.g., migraines, lower back pain). Psychiatry is unique in the sense that this problem faces the whole field because of the nature of mental disorder and the way it is defined. On the one hand, we believe that patients with psychiatric conditions may be entitled to a range of benefits allocated to those who are deemed to be sick. On the other hand, is individual psychiatrist appraisal a satisfactory standard by which to make these judgments?

Psychiatric expertise

A second challenge facing the field of psychiatry is to know exactly where psychiatric expertise lies. Because psychiatry exists within the medical system it must adhere to its parameters. It would make sense then that psychiatrists, like other physicians, would locate their expertise in diagnosis and treatment. However, diagnosis in psychiatry lies in applying *DSM* criteria, a task that non-medical clinicians such as psychologists also undertake, a task that family physicians undertake much more often than psychiatrists, and one that can be achieved through the administration of structured interviews (whether by physician, non-physician, or trained non-clinician) rather than necessarily by a psychiatrist (First et al., 2016: 1). Unlike in other areas of medicine it is not clear that diagnosis is the central area of expertise for psychiatrists.

What about treatment? Is there a clearly defined role for psychiatric expertise in the realm of therapeutic intervention? There are a range of interventions—pharmacological, psychotherapeutic, social or rehabilitative—available for mental disorders, depending on the specifics of the problem. Psychiatrists are trained in the use of pharmacological treatments but their ability to offer other types of treatments will vary. As a professional group their expertise does not lie in the mastery of non-pharmacological treatments (although individual

psychiatrists may well be expert in the delivery of these treatments). Yet there are important problems in terms of professional identity if psychiatric expertise is psychopharmacological expertise. Psychiatric medications are classed by their actions on symptom clusters but even then, they are prescribed for ever-widening indications. Thus, they are quite non-specific. In addition, there are important debates surrounding the effectiveness (Kirsch et al., 2008; Turner et al., 2008; Vohringer and Ghaemi, 2011) and the harmfulness (Moncrief, 2013; Wade et al., 2017) of many existing psychiatric medications. If the core of psychiatric expertise lies in the prescription of these medications it makes psychiatrists experts of a dubious therapeutic armamentarium. Bluntly put, psychiatrists would be experts in the use of treatments with questionable therapeutic benefit at best, harmful at worst.

Psychiatrists face the challenge of how to establish their disciplinary expertise within the medical profession and the medical system. Medical care naturally focuses on those aspects of patients' problems that conform to the medical model; that is, the biological component. While this may make sense for a condition like pneumonia where the pathophysiological process is the largest contributor to the patient's problem it seems less evident that this should be the case with mental disorders where biological disease may play a smaller role in a given clinical problem. Should psychiatrists confine their roles to those problems or parts of problems which are or seem to be biological? Current trends in psychiatric training, practice, and research seem to be oriented in this direction. This makes the fit between psychiatry and medicine more coherent but leaves open the question of who should respond to the psychological and social dimensions of mental disorders and how such services should be organized and financed.

Conclusions

Psychiatry operates within the institutional structures of medicine and shares some of its conceptual and practical characteristics. At the same time, it shares methods with other non-medical clinical disciplines. Despite this unique interdisciplinarity, the field of psychiatry faces numerous issues, including uncertainty about the nature of mental disorder, controversy about treatments, and competition with other disciplines. The continued lack of clear biological disease undermines its place in medicine while opening it up to non-medical challengers who claim to do an equally good job. Policymakers take note of the fact that much of what psychiatrists do—diagnostic evaluation, prescribing

psychotropic medication, psychotherapy—is also being done by family physicians or other clinicians whose salaries are lower and whose professional lives are more amenable to control. Meanwhile, the role for which psychiatrists seem irreplaceable is their medicolegal one: as arbiters of a host of thorny ethical questions such as which patients should be compelled to received hospitalization or treatment, are criminally responsible for their acts, or are entitled to receive a variety of social benefits. Is this the vision that psychiatrists have of their work? Is this the version of psychiatry that patients and family members need? Despite trends to improve the science of psychiatry, as witnessed by the influence of evidence-based practice approach, the use of standardized diagnostic and symptom rating scales, and the manualization of psychotherapy, it is when psychiatrists are asked to make ethical judgments about mental life that the distinct character of the field of psychiatry becomes clear.

Notes

1 Exceptions include in certain jurisdictions, the prescribing powers of nurse-practitioners or of pharmacists to prolong or initiate prescriptions.
2 In certain situations, such as the lack of availability of a psychiatrist, the authorization of a general physician can be sought.

References

American Board of Psychiatry & Neurology (2011), Psychiatry Core Competencies Outline. URL= https://www.abpn.com/access-residency-info/residency-training-information/psychiatry/ Retrieved April 26, 2017.

American Psychiatric Association (2013), *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Arlington, VA: American Psychiatric Publishing.

Engel, G. L. (1977), The Need for a New Medical Model: A Challenge for Biomedicine, *Science*, 196:129–136.

First, M. B., Williams, J. B. W., Karg, R. S., and Spitzer, R. L. (2016), *User's Guide for the Structured Clinical Interview for DSM-5 Disorders-Clinician Version (SCID-5-CV)*, Arlington, VA: American Psychiatric Association.

Ghaemi, S. N. (2009), The Rise and Fall of the Biopsychosocial Model, *The British Journal of Psychiatry*, 195:3–4.

Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., and Johnson, B. T. (2008), Initial Severity and Antidepressant Benefits: A Meta-Analysis of Data Submitted to the Food and Drug Administration, *PLoS Medicine*, 5:260–268.

Moncrieff, J. (2013), Long-Term Antipsychotics—Making Sense of the Evidence. URL= https://joannamoncrieff.com/2013/12/09/long-term-antipsychotics-making-sense-of-the-evidence/ Retrieved April 26, 2017.

Murphy, D. (2011), Conceptual Foundations of Biological Psychiatry, pp. 42–427. In *Handbook of Philosophy of Science Volume 16 Philosophy of Medicine*. F. Gifford, ed. London: Elsevier.

Royal College of Physicians and Surgeons of Canada (2015), *Objectives of Training in the Specialty of Psychiatry*. Version 2.0. URL= http://www.royalcollege.ca/rcsite/education-strategy-accreditation/pgme-training-programs/accreditation-residency-programs-e Retrieved April 26, 2017.

Sadock, B. J. and Sadock, V. A. (2011), *Kaplan and Sadock's Synopsis of Psychiatry*, 10th edition. Philadelphia, PA: Williams & Wilkins.

Schwartz, S. J., Lilienfeld, S. O., Meca, A., and Sauvigné, K. C. (2016), The Role of Neuroscience within Psychology: A Call for Inclusiveness over Exclusiveness, *American Psychologist*, 71:52–70.

Sturney, P. (2009), Case Formulation: A Review and Overview of This Volume, pp. 3–30. In *Clinical Case Formulation: Varieties of Approaches*. P. Sturney, ed. Chichester: Wiley-Blackwell.

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., and Rosenthal, R. (2008), Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy, *New England Journal of Medicine*, 358:252–260.

Vöhringer, P. A. and Ghaemi, S. N. (2011), Solving the Antidepressant Efficacy Question? Effect Sizes in Major Depressive Disorder, *Clinical Therapeutics*, 33:B49–B61.

Wade, M., Tai, S., Awenat, Y., and Haddock, G. (2017), A Systematic Review of Service-User Reasons for Adherence and Nonadherence to Neuroleptic Medication in Psychosis, *Clinical Psychology Review*, 51:75–95.

Study Questions

1. *What is philosophy of psychiatry?*
2. *What is the difference between psychiatry and other mental health professions?*

Part Two

Philosophy of Mind and Psychiatry

3

Understanding the Nature of Mental States: Psychiatry, the Mind-Body Problem, and the Biopsychosocial Model of Medicine

Jesse Butler

Introduction

Given that psychiatry is the area of medicine devoted to mental disorders, it is crucial for psychiatry to have a viable understanding of the nature of mental states. However, the nature of mental states has been a persistent source of controversy throughout history to the present day, with no clear agreement in sight among the contemporary philosophers and scientists engaged in the topic. This puts psychiatry in a difficult position. If psychiatrists are to diagnose and treat mental disorders, the lack of agreement concerning our basic understanding of mental states appears to be a significant barrier to effective procedures and results.

Fortunately, there is a way to situate psychiatry in a functional position here despite the lack of agreement across viewpoints concerning its basic domain. Rather than enlisting in the battle over the nature of mental states, seeking to nail down a single theory or explanation for itself, psychiatry can instead take advantage of the plurality of available viewpoints and apply them as best fits its aims. The goal of this chapter is to establish a philosophical basis for this approach in psychiatry, utilizing what is known as the biopsychosocial (BPS) model of medicine. As its name suggests, the BPS model is a pluralistic approach to disease and health that integrates multiple modes of explanation, diagnosis, and treatment, rather than solely privileging any one perspective over others. To be clear, this does not mean that "anything goes," such that any perspective or theory of mental states is as good as another. Rather, it is the recognition that mental states and their various properties are exceedingly complex and thus should be approached from multiple angles rather than just one, with those

various perspectives still being held to reasonable standards of accuracy and demonstrable success.

To make a case for the BPS model in psychiatry, I will first survey the main philosophical views of the mind, taking a look at what is known as the mind-body problem in philosophy. This will enable us to see the main contending options for understanding the nature of mental states, as things stand today in both philosophy and the sciences of the mind. As we will see, much of the debate comes down to whether or not mental states are brain states. I will then argue that even if mental states are brain states, this does not privilege one single mode of explanation over others. Instead, following the BPS model, brain states exist within an overlapping matrix of underlying mechanisms, cognitive functions, and social behaviors, such that these factors must be integrated together to effectively understand and address mental disorders in psychiatry.

The mind-body problem

It is common to begin analyses of the mind-body problem with the seventeenth-century French philosopher René Descartes, who advocated a distinct separation of mind from body. Descartes was not the first person to think that the mind is distinct from the body, of course, but his articulation of mind-body dualism set the stage for much of modern philosophy. According to Descartes, and the doctrine now often known as Cartesian Dualism, a mind is a non-physical substance, more commonly referred to as a soul that is distinctly separable from a physical body and may exist without it. For example, in a famous passage from his *Discourse on Method*, Descartes states:

> Then I looked carefully into what I was. I saw that while I could pretend that I had no body and that there was no world and no place for me to be in, I still couldn't pretend that I didn't exist … This taught me that I was a substance whose whole essence or nature is simply to think and which doesn't need any place, or depend on any material thing, in order to exist. Accordingly this me—this soul that makes me what I am—is entirely distinct from the body, is easier to know than the body, and would still be just what it is even if the body didn't exist. (Descartes 1637, Part IV)

If this perspective is correct, then mental states are aspects of an immaterial soul, the health of which would fall under the purview of religion and spirituality rather than the earthly physiological concerns of modern medicine. However, few theorists of the mind subscribe to this kind of dualism today, especially

in light of the subsequent scientific developments following the early modern period of Western philosophy. For example, in an introduction to an anthology of contemporary perspectives in the philosophy of mind, James Garvey writes:

> Perhaps Descartes could take seriously the notion of a soul existing apart from the body—an unextended thing located precisely nowhere yet somehow running the show—but we no longer can. After all, Descartes grew up in a world which had only just started to assimilate the notion that Earth was not the centre of the universe. As our understanding of such things as biology, evolution, psychiatry and physics expanded, the old view of the mind just didn't fit with the rest. How do you bolt a soul onto an evolved organism? What use is a thinking substance to theorizing about our psychological states? Where do you put soul stuff when your ontological inventory seems complete with matter and energy? (Garvey 2015: xx)

Despite widespread rejection of mind-body dualism, however, the relationship between mind and body continues to be a matter of significant debate. In fact, this quote from Garvey appears in an introduction intended not simply to display contemporary rejection of Cartesian dualism but to lay a foundation for debate across the various significant theories and issues in one of the most active areas of philosophy today. We cannot cover all of that territory here, but an overview of some of the main options will be instructive.

With the rejection of unobservable Cartesian minds, some influential philosophers (e.g., Ryle 1949) and scientists (e.g., Skinner 1953) first turned to observable human behavior to understand the nature of mental states. There are different kinds of behaviorism, but they all share the attempt to relate mental states to behaviors and behavioral propensities. For example, fear is understood in terms of the tendency to avoid a given object or situation, while desire is the tendency to pursue a given object or situation. Behaviorism played a significant role in establishing psychology as a science, and in philosophy moving beyond Cartesian dualism. Indeed, its influence is still prominent today, given that mental states are often understood in relation to behaviors. In psychiatry, this is evident in the emphasis on observable patterns of behavior as symptoms of mental disorders, as can be seen across the criteria in the current diagnostic manual for mental disorders, the *DSM*-5 (American Psychiatric Association 2013). We will see this in action when we consider the case of major depression below, such that behaviors like "appearing tearful" are utilized in making diagnoses.

However, few theorists regard themselves as behaviorists today, given that behaviorism faces some devastating failures as an adequate approach to mental states. In retrospect, the deficit is not hard to see, as revealed in this classic joke about behaviorism:

What do behaviorists say after sex?

"That was good for you, but how was it for me?"

A narrow focus on observable behavior leaves out what is most interesting and important when it comes to understanding mental states: our own first personal thoughts and experiences. While behaviors relate to mental states in various ways, most agree that mental states themselves cannot be strictly identified with behaviors alone. Mental states can cause behaviors (and vice versa), but are not behaviors themselves. In psychiatry, this point can be seen in the recent shift from behavioral symptoms as observable markers of mental disorders, as highlighted in the *DSM*, to understanding the underlying causes that produce those symptoms, as proposed by the National Institute for Mental Health's Research Domain Criteria (RDoC), a new research paradigm aimed at advancing psychiatry beyond the symptom-based approach of the *DSM* (Cuthbert and Insel 2013).

If mental states are neither observable behaviors nor unobservable properties of an immaterial soul, however, then what exactly are they? As you are likely aware, the most prevalent viewpoint today is that mental states are brain states. A fear of spiders might reliably correlate with a behavioral tendency to scream and run upon seeing a spider, but the fear itself is contained in the fearful person's head, encoded somehow in the network of neural processes embodied in the brain. This brings us to what is known as the identity theory, meaning that mental states are identical to brain states (Smart 2017). There are multiple versions of the identity theory, but one particularly reductionist formulation of the theory is summarized nicely in Francis Crick's memorable proclamation that "you're nothing but a pack of neurons" (1994: 3). Applied to psychiatry, the implication is that mental disorders are brain disorders, often assumed to be best understood through neuroscience and treated through biomedical procedures applied to the brain (e.g., pharmaceutical drugs administered to change brain chemistry).

Yet again, however, this is not the end of debate, but rather a new beginning. Consider this question: How exactly do brains establish thoughts and experiences, such that mental states and brain states can be explained together as the same thing? This brings us to the heart of controversy in the study of the mind today. We cannot simply open up a person's head and see a fear of spiders or a state of depression in the gray matter within. Moreover, it is generally agreed that a dead brain, while still a brain, does not have mental states. Considering these and other related points, the focus of inquiry has turned to understanding what

brains do: What exactly are the neurons and neurotransmitters doing inside our heads, to give rise to the thoughts and experiences we know as our mental states through our own first person consciousness?

One leading approach to this issue, which has laid a foundation for the contemporary interdisciplinary field of cognitive science, is to understand the brain in terms of cognitive functions. Known as functionalism in the philosophy of mind, this theory proposes that mental states are best understood in terms of the informational functions they perform, rather than the underlying mechanisms that embody those functions in the brain (Levin 2016). To use a common computer analogy, mental states are the software running through the operations of an actively functioning brain, rather than the underlying hardware itself. Just as computer hardware is useless without software installed to perform various functions, so too is a brain simply a mindless mass without some kind of active functional organization that gives rise to the informational processes that constitute the mental states it embodies. There are different forms of functionalism, varying in terms of the kinds of functions they attribute to mental states, but they all share the idea that mental states are best approached in terms of cognitive functions performed in the processing of information. Applied to psychiatry, functionalism entails that mental disorders can be understood as problems with the cognitive processes embodied in the brain, as emphasized in the relatively new and growing field of cognitive neuropsychiatry (Halligan and David 2001). Just as solving a computer problem might involve detecting and fixing a software glitch, without necessarily changing any of the hardware, so too might psychiatry function not by intervening at the hardware level of the brain but rather by addressing the cognitive states of a patient, say a delusional belief or irrational fear, for instance, and utilizing cognitive therapy to treat the problem.

Despite its prominence in the philosophy and sciences of the mind, however, functionalism faces a number of criticisms, from multiple angles. Some influential contemporary philosophers argue that functionalism, like its predecessor behaviorism, cannot explain the most interesting and important feature of mental states, namely, consciousness. This critique leads some thinkers to advocate a form of dualism known as property dualism (e.g., Chalmers 1996), treating conscious mental states as non-physical properties that may emerge from the brain but are not reducible to functional or physical processes in the brain. To be clear, property dualism, as a contending theory in contemporary philosophy of mind, is quite different from Cartesian dualism. Property dualists do not posit a distinct non-physical soul as the seat of consciousness, and

generally allow that conscious mental states are dependent upon and/or causally connected to the brain. Their point is that a reductive focus on underlying causes of mental states in the brain does not capture everything there is to know about them.

Mainstream functionalism, with its focus on rather high-level cognitive functions (e.g., memory, attention, language, and so on), faces objections from more reductive thinkers as well. Some theorists argue that we will not fully understand mental states until we can explain them in terms of their underlying biological mechanisms, treating functionalism as just an incomplete phase in the process of scientific explanation, rather than a sufficient theoretical framework itself. From this line of criticism, mainstream forms of functionalism operate at a deceptively overgeneralized level, failing to grasp the actual causal mechanisms at work in the brain. Theorists of this sort argue that we should aim to reduce, or even eliminate, our coarse-grained functionalist understanding of mental states in favor of more fine-grained explanations of what is actually going on in the brain (Bickle 2003; Churchland 2005). From this perspective, we may not be able to address mental disorders effectively until we can reduce them to their underlying physiological causes in the brain, giving priority to fine-grained understanding and treatment at the level of mature neuroscience, which may itself be reduced to even more basic mechanisms. Identifying a genetic condition that gives rise to chemical dysfunction in the brain, for instance, would be a mark of success from this standpoint, enabling discovery of its underlying causal mechanisms rather than its behavioral symptoms or higher-level cognitive effects.

So, while there has been a clear convergence toward the brain in our understanding of mental states, significant debate remains and the classical mind-body problem is still very much alive today, not just as an intellectual problem, but also a practical problem with significant implications for psychiatry. As we have seen, disagreement largely revolves around which level of analysis we should take in understanding mental states. Can mental states be understood in terms of behavioral symptoms, such that observing behavioral patterns is sufficient for diagnosing and treating mental disorders? Should we instead focus on the mediating cognitive functions at work in human psychology, such that mental dysfunctions can be treated at the cognitive level? Or must we look at the fine-grained electrochemical processes embodied in the brain to truly understand and treat mental disorders? Alternatively, could the prioritization of these various objective features somehow neglect the importance of the patient's own lived experience and the relevance of their first person consciousness in understanding a mental issue they face?

The biopsychosocial model

Given the range of debate concerning mental states, what should psychiatry do? The suggestion I offer here is that psychiatry need not take a strong stance in privileging any one viewpoint or level of analysis, but rather can utilize the plurality of perspectives concerning the nature of mental states. In terms of the viewpoints surveyed above, behaviors, neural mechanisms, cognitive processes, and conscious experiences can all be relevant to the understanding and treatment of mental disorders, with psychiatry being able to benefit from the diverse considerations that these various facets of mental states offer. We humans like to have a nice single theory to explain things, and often lean toward one pet theory or another, but in the case of understanding mental states and disorders, they are so complicated and multifaceted that we should not expect any single theory alone to be sufficient. Instead, a pluralist acceptance of multiple theories and levels of analysis is the most plausible route to take in confronting the nature of mental states and the complexities of mental disorder, as experienced by whole persons within the context of their lived existence in the world.

In order to explain and defend this pluralist approach to psychiatry, I will draw upon the BPS model of medicine. Although earlier precedents for this model may be identified, it is commonly associated with an influential and widely cited article by George Engel, "The Need for a New Medical Model: A Challenge for Biomedicine" (1977). In this article, Engel criticizes the dominant biomedical model, claiming that it is a relic of outdated mind-body dualism (such as we saw earlier with Descartes), and offers a new BPS model as a more scientifically accurate and advantageous approach to medicine. In what follows, I will overview Engel's BPS model and illustrate how it offers a viable pluralistic understanding of mental states and mental disorders through the example of depression.

To begin, we must first contrast the BPS model with Engel's object of criticism: the biomedical model. From this perspective, the human body is seen as a biological machine, with disorders in that machine being treated through biomedical intervention, without regard to other facets of a person's life, such as their personal thoughts and experiences or interpersonal relationships. According to Engel's critique, this isolation of the human body from other aspects of a person's life is a product of mind-body dualism in Western Christian culture:

> With mind-body dualism firmly established under the imprimatur of the Church, classical science readily fostered the notion of the body as a machine, of disease as the consequence of breakdown of the machine, and of the doctor's task as repair of the machine. (p. 131)

The other aspects of the person, their thoughts and experiences, behaviors and social relationships, were thus conceived to be outside the purview of medicine, falling instead within the religious domain of care of the soul. The soul being distinguished from the body, the body was left to be the object of mindless mechanical analysis, explanation, and treatment. Mind-body dualism thus led to biological reduction in medicine, with the ill body reducible to purely mechanical dysfunction.

What is the problem here? The conception of the human body as a biological machine has led to tremendous medical advancement, such that we can now treat a wide variety of medical conditions through biomedical interventions with surgery, pharmaceutical drugs, and so on, so why shouldn't we embrace this biomedical model? First of all, like most contemporary theorists, Engel rejects mind-body dualism as a problematic and inaccurate foundation for understanding human life. Engel sees the behavioral, psychological, and social aspects of human existence as part of the natural world, falling within the scope of medicine rather than outside of it. Medicine thus should not be treated as pertaining solely to biological mechanisms, but rather to the whole person, across all facets of human life, including, but not limited to, biological mechanisms.

Secondly, the separation of the mind from the body puts psychiatry in a tenuous position, such that it must be either excluded from medicine or reduced to solely biological parameters. With the exclusionary option, psychiatry itself would be ill-conceived as a field of medicine, and thus rejected insofar as it is intended to address mental illness as a medical phenomenon. On the other hand, if psychiatry is conceived as a field of medicine under the biomedical model, it would be confined to merely biological considerations, thereby automatically reducing all "real" mental illness to mechanical dysfunction in the body. Engel saw both of these options as deeply problematic, leading to a state of crisis in psychiatry by forcing it between two undesirable options. Engel's solution is to posit a new model of medicine that avoids the problems of dualism and biological reductionism, reframing medicine in a way that integrates, rather than ostracizes, the mental domain of human life.

As the name itself reveals, the BPS model advocates the consideration of biological, psychological, and social factors together in the understanding and treatment of disease. Engel did not seek to reject the biological dimension emphasized by the biomedical approach, but rather to also include psychological and social dimensions in the central considerations of medicine, thereby healing the erroneous divide between these factors brought about by mind-body dualism. In other words, the whole person falls within the proper purview of

medicine, rather than simply reducing a person to the biological processes of the human body. Although we will focus on its application to psychiatry, it is worth noting here that Engel intended the BPS model to apply to all of medicine, for he thought that it would be both more humane and scientifically viable to include the whole person in doctors' interactions with patients and in the general understanding and treatment of disease.

Engel developed the BPS model in accordance with the general systems theory of the Austrian biologist Ludwig von Bertalanffy (1968), utilizing this theory as an integrative framework for connecting disparate levels of science together, across different levels of organization. Ranging from particle physics to the biosphere of the Earth, human persons are situated within a hierarchical arrangement of organizational systems, including biochemical processes, cellular structures, organ systems, the whole person as a human organism, the person's family and social relations, and their broader fit within an environmental ecology (Engel 1977: 196):

Biosphere

Society

Family

Individual Person

Organ Systems

Cellular Structures

Molecules

Although the BPS model is not necessarily reliant upon the particular details of general systems theory, the arrangement of different levels of organization into an inter-relational system illustrates how the biological, psychological, and social facets of a person can be brought together into a unified yet pluralistic understanding, rather than competing with each other for explanatory priority.

Returning to the seemingly conflicting theories of mental states surveyed earlier, we can reconceive the various contending accounts of mental states as interconnected facets of a human person. With the brain as the organic seat of mentality within the human organism, it is causally connected to the observable behaviors exhibited by the person within the context of their social life, and itself composed of underlying biomechanical processes embodied in the complex array of neurons and their electrochemical activity. These

processes, in turn, can be understood in terms of the cognitive functions they perform, situating the software-level analysis of cognitive science between the behaviors of the whole person and the underlying biological components of the brain. Through the BPS model, we can see that these levels of analysis are not competing accounts of mentality, but rather integrated facets of a person as viewed by higher and lower levels of analysis. While reductive biological explanation provides tremendous insight, it is more insightful and productive when also situated within a more encompassing framework. As further integrative work in psychiatry has confirmed, "Through recent research on genetics and neurobiology, it has become increasingly clear that the social combines with, and impacts on, the biological and the psychological, in a multitude of ways, to increase risk for mental disorders" (Reininghaus and Morgan 2014: 1).

Moreover, conscious thoughts and experiences provide yet another level of organization and analysis, situating the first person perspective within the broader matrix of biological mechanisms, psychological functions, and social behaviors. Property dualists treat this level of organization as irreducible to the others, arguing that consciousness is not amenable to physical explanation. This might seem to be a barrier to the natural unity of the BPS model, however property dualists generally accept at least some kind of causal relationship between mental states and brain states, which still connects them to the naturalistic processes within the scope of the BPS model. In fact, David Chalmers, a leading property dualist, describes his view as naturalistic dualism, such that conscious properties, although not physical, are nevertheless a part of the natural world (1996). Additionally, many contemporary philosophers argue that the first person character of consciousness can be accommodated within the natural world without positing anything non-physical. As one example, I refer interested readers to my book on introspection, where I argue that our experiential knowledge of our own conscious mental states exists in virtue of our embodied existence as natural beings in the world, eliminating the need to posit a non-physical domain to account for the unique experiential character of consciousness (Butler 2013).

To illustrate the implications of the BPS model in practice, let us consider a case of clinical depression. Suppose Bill is visiting his doctor with complaints that he has had trouble getting to sleep for the past few months. His physician inquires further, and after a survey of Bill's symptoms, including, let us suppose, confirmation from Bill's wife and elimination of other possible medical conditions symptomatically related to depression, the doctor determines that

Bill meets the following criteria from the *DSM*-5 for major depressive disorder (American Psychiatric Association 2013: 198–199):

1. Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad, empty, hopeless) or observation made by others (e.g., appears tearful).
2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation).
3. Insomnia or hypersomnia nearly every day.
4. Fatigue or loss of energy nearly every day.
5. Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).

On the basis of these five collectively sufficient criteria, Bill's doctor diagnoses him with major depressive disorder and writes a prescription for an antidepressant, telling him that it should help alleviate his symptoms, and schedules a follow-up meeting to later confirm whether that is the result.

Although biological, psychological, and social factors enter into this process, it would be remiss to see the resulting diagnosis here as an example of following the BPS model. Instead, what has happened is that these various factors have been automatically funneled into a default biomedical approach, reduced to a presumed biological cause to be treated through medical intervention in the body, without further addressing the behavioral, psychological, or social aspects of Bill's life.

Now, it may be that Bill could benefit from this result, if, for instance, he is suffering solely from a serotonin imbalance that can be alleviated by an SSRI antidepressant. However, from the information given, it is not at all clear that this is the case despite the fact that this kind of process and result is quite common in standard medical practice (Horwitz and Wakefield 2007). Notice that, while various features of Bill's mental states are considered, they are not considered with further systematic attention to their context in Bill's life. Furthermore, only one biologically reductive causal explanation is considered, through rather vague inference from behavioral markers to presumed neurochemical dysfunction (rather than, say, direct measurement and assessment of neurochemical health, which science itself has not yet progressed to the point that such an approach can be implemented).

To see the deficits with this process and its assumptions, suppose further that Bill is recently retired from a long and stable career that played a large role

in giving his life meaning and structure. The non-contextual symptomology of the *DSM* criteria alone could easily allow such a fact to go unnoticed, but this contextual fact about Bill's life could be tremendously important in understanding the cause and prospective treatment of Bill's depression (Horwitz and Wakefield 2007). Even if we follow a reductive identity theory here and regard Bill himself as just a pack of neurons, it is important to understand what the broader functional or dysfunctional state of those neurons may be, in relation to his behaviors, relationships, and social circumstances. Neurons do not exist in a vacuum. If Bill's depressive changes in neurochemistry correspond with dramatic changes in his daily behavioral habits and cognitive processes, following his recent retirement, then those features of Bill's life are relevant to understanding and therapeutically addressing the pack of neurons in his head. While an SSRI may help alleviate the depressive emotional affect of Bill's brain, it may not truly address the central causal factor behind Bill's depression or help him establish more functional thoughts and behaviors in his life. Causality and therapeutic intervention can operate on multiple levels, and should not be conflated with biological mechanism alone (Kendler and Campbell 2009). Perhaps Bill would be better served by some form of cognitive-behavioral therapy or life counseling to help him adjust his thoughts and behaviors to his new phase of life in retirement, and a medical practitioner following a pluralistic understanding of their patients' lives through a BPS model would be more likely to consider and prescribe such options.

Although this is just one example, these considerations reveal the significance of considering the BPS model in psychiatric understanding of mental states, across the symptoms, causes, and treatments of mental disorders. While mental states arguably are brain states, they are also causally connected to behavior and perform psychological functions operating within the context of human social life. The BPS model prompts psychiatry to address these factors together, as a framework for integrating these complex and interrelated facets of mental states into our understanding of the human mind and applying that understanding to the diagnosis and treatment of mental disorders.

Objections and replies

Despite continued dominance of the biomedical model, the BPS model has achieved notable influence and should continue to develop as psychiatry advances. With the emergence of evolutionary psychiatry, for example, there

are new possibilities for further integrating the BPS facets of mental disorders (Nesse 2005). Nevertheless, the BPS model has met with significant criticism as well. In this concluding section, I will confront some main objections against the BPS model and its implications for understanding mental states in psychiatry.

Although it is intended to overcome problems with mind-body dualism, some critics of the BPS model claim that the model itself generates a problematic divide between the biological and the psychological. For example, Tavakoli states:

> Unfortunately, for most novel users attempting to learn and use the biopsychosocial model, the model may suggest that biology is separate from psychology. While trying to teach future psychiatrists about major depressive disorder, for example, how can we, as educators, delineate for them what is biological versus what is psychological? When the biopsychosocial model is recommended as a teaching model in the halls of academic medicine, one can see the frustration in medical students or junior residents trying to categorize a patient's symptoms and clinical history suggestive of major depressive disorder into biological versus psychological. The ambiguity of such distinctions highlights the weakness of this model for teaching and clinical purposes. (Tavakoli 2009: 26)

This is an unfortunate situation, but not because of the BPS model itself. Instead, the problem is a misunderstanding of the model. As we have seen, a fundamental goal of the BPS model is to bring together multiple facets of mental states into an integrated whole. Biological mechanisms, psychological processes, and social behaviors all hang together as features of a whole person and their experiences in life. These are not discrete and separable elements, to be parsed apart by medical practitioners in deciding whether a case of depression is biological or psychological. Depression is both biological and psychological at the same time. The purpose of the BPS model in medical practice is to consider these factors together, as they pertain to the symptoms and circumstances of a particular patient. Considering this, Tavakoli's critique here should not be aimed at the BPS model itself, but rather its erroneous portrayal and application in psychiatry.

Some critics, however, have questioned whether the BPS model actually creates an objectively viable approach across the diverse phenomena it encompasses. For example, perhaps the most outspoken critic of the BPS model, Nassir Ghaemi, argues that the BPS model fosters a relativistic "anything goes" eclecticism that he associates with postmodern loss of objective truth (2009). In a recent book on depression, he summarizes his critique as follows:

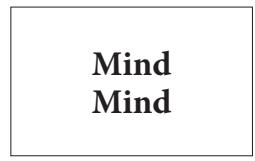
> What started with some viable rationale was soon infected with the postmodern virus. The infection began when clinicians began to ask the question: Okay, given that all illness has biological, psychological, and social components, which should we emphasize in this case, for this patient, in these circumstances, with this condition? The BPS model provided no general answer to specific clinical questions. It was like saying the law of gravity exists but never being able to say how it specifically applied in the world to a physical object. Or, as some early critics suggested, it was like going to a restaurant to eat, and, instead of a menu, receiving a list of ingredients. One might admit that there are biological, psychological, and social ingredients to many illnesses, but in what combination? Are they always equally influential? Or should one prioritize one versus the other at times? The BPS model had no ready answer, so clinicians improvised. Their answer was eclecticism: anything goes; nothing is forbidden. I as a clinician may decide to emphasize psychotherapy; you as a patient may decide to focus on medications. Everyone is free to do whatever anyone wants to do, and there is no wrong answer. All answers are right. (Ghaemi, 2012: 59)

To address these concerns, a few different points must be made. First of all, it is crucial to distinguish between pluralism and relativism here. Pluralism, as I have used the term, is the view that multiple perspectives, considerations, and/or theoretical frameworks are relevant to a particular topic or issue. So, as we have seen, the idea that biological, psychological, and social factors pertain to understanding mental states is a kind of pluralism. Relativism, on the other hand, is the view that truth itself is relative to one's perspective, culture, or theoretical approach. As we see in the quote from Ghaemi, this rejection of objective truth is often associated with postmodernism and its critical rejection of foundational objective truth in philosophy and science. However, these concepts should not be conflated with each other. Pluralism itself is not a rejection of objective truth, and the fact that multiple perspectives or theories may apply to a given phenomenon simply does not entail that the truth of the matter is relative to whichever perspective or theory you pick, such that "anything goes" and people are free to say and do whatever they wish.

Consider the case of Bill again. If we attend to his neurological state, psychological processing, social behaviors, and experiential emotions, we are not forced to choose which of these are true, nor is the truth relative to one of these factors alone. If it is true that his depression is a result of his recent retirement, for instance, this does not negate the fact that he may exhibit neurochemical changes as well. These things can be true together, if we regard them as multiple facets of one and the same thing in reality: the

human person Bill himself. Additionally, if we are comparing pharmaceutical intervention with cognitive-behavioral therapy as treatments for Bill, the truth of which is most appropriate is not a relativistic choice of personal preference but rather a pragmatic choice regarding what would best address the problems Bill is facing. Perhaps a mix of both would be appropriate, depending on the circumstances. Considered in practice, the BPS model is not a postmodern free-for-all, devoid of objective truth, but rather a nuanced and practical consideration of the multiple real factors pertaining to human mental life.

To further clarify how the BPS model can apply diverse theoretical perspectives to particular individuals and their mental states, it will be helpful to introduce a standard philosophical distinction between types and tokens (Wetzel 2014). A token is a particular concrete instance of a thing, while a type is a general theory-specific category that may be applied to a thing. To illustrate, consider the following question: How many words are in this box?



The question itself is confusing because the number of words is ambiguous, but the type/token distinction dissolves the problem: There are two tokens of one type of word. At the token level, the word "mind" appears twice in the box, so there are two concrete instances, or tokens, of the word. But both of these word tokens are of the same type, both being in the same general form. Notice that objective truth has not been forsaken here, despite the fact that we began with an ambiguous question and now have two answers. It is an objective fact that there are two tokens of one word type.

Now consider this question: How many things are in the following box?



Here again we have an ambiguous question, which must be clarified by asking what type of thing we are counting. If we are counting tokens of the word "things," then the answer is clearly one. But if we are counting letters, then the answer would be six. Does this mean that we have fallen down a postmodern relativist wormhole where anything goes, and anyone can say there are however many things they want? No. It just means that we have different types of categories to consider, each of which can have clear objective facts within its proper domain.

Applying the type/token distinction to mental states, any particular token mental state can be identified as a single token physical state, such that it is understood to be a physical event in a brain. At the same time, however, that individual token state can be described by a variety of types, differential to whatever theoretical structures are applicable in a given context, for example understanding the state in terms of the electrochemical activity studied by neurology, the cognitive functions studied by psychology, and the behaviors studied by the social sciences. Each of these distinct theory-based types may apply to what we can accept as one and the same token thing: the mental state of a physical human being.

As one last analogy, suppose a geologist is investigating the age of the Grand Canyon, utilizing geochemistry to study rock composition, fluid dynamics to study the impact of the Colorado River on the rock, and plate tectonics to study large-scale shifts of land. Just as these various theoretical perspectives may converge toward objective understanding of a single token object of study, without necessary conflict or competition, so too may a psychiatrist investigate a particular mental disorder through multiple theoretical frameworks that are all applicable to one and the same token subject: the mental life of a human being. In both cases, multiple considerations can operate together in an accurate and integrated manner, without dismissing objective truth or devolving into relativistic eclecticism.

Through the clarity provided by the type/token distinction, we can see that the BPS model is not impeded by the presence of multiple theories of mental states. Instead, we must accept that the human mind is a complex phenomenon, amenable to a variety of explanatory frameworks. The BPS model thus should not be expected to provide a single theory to utilize across all psychiatric practice. Indeed, it itself is not an explanatory theory, but rather a pluralistic framework for integrating multiple theories in application to the complex and multifaceted domain of mental illness and health, ultimately rooted in the objective reality of human beings in the world. Thus, the unity of the BPS model is in the reality

of concrete human life itself, regarding whole persons and the complex web of factors pertaining to their lives. In this respect, the broad encompassing nature of the BPS model is a virtue rather than a drawback.

References

American Psychiatric Association (2013), *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Arlington, VA: American Psychiatric Publishing.

Bertalanffy, L. von (1968), *General Systems Theory*, New York: Braziller.

Bickle, J. (2003), *Philosophy and Neuroscience: A Ruthlessly Reductive Account*, Dordrecht: Kluwer.

Butler, J. (2013), *Rethinking Introspection: A Pluralist Approach to the First-Person Perspective*, New York: Palgrave Macmillan.

Chalmers, D. J. (1996), *The Conscious Mind: In Search of a Fundamental Theory*, New York: Oxford University Press.

Churchland, P. M. (2005), Functionalism at Forty, *Journal of Philosophy*, 102:33–50.

Crick, F. (1994), *The Astonishing Hypothesis: The Scientific Search for the Soul*, New York: Touchstone.

Cuthbert, B. N. and Insel, T. R. (2013), Toward the Future of Psychiatric Diagnosis: The Seven Pillars of RDoC, *BMC Medicine*, 11:126.

Descartes, R. (1637), *Discourse on the Method of Rightly Conducting One's Reason and Seeking Truth in the Sciences*. J. Bennett, trans. 2007. URL= www.earlymoderntexts.com

Engel, G. (1977), The Need for a New Medical Model: A Challenge for Biomedicine, *Science*, 196:129–136.

Garvey, J., ed. (2015), *The Bloomsbury Companion to Philosophy of Mind*, New York: Bloomsbury Academic.

Ghaemi, N. (2009), *The Rise and Fall of the Biopsychosocial Model: Reconciling Art and Science in Psychiatry*, Baltimore, MD: Johns Hopkins University Press.

Ghaemi, N. (2012), *On Depression: Drugs, Diagnosis, and Despair in the Modern World*, Baltimore, MD: Johns Hopkins University Press.

Halligan, P. W. and David, A. S. (2001), Cognitive Neuropsychiatry: Towards a Scientific Psychopathology, *Nature Reviews Neuroscience*, 2:209–215.

Horwitz, A. V. and Wakefield, J. C. (2007), *The Loss of Sadness: How Psychiatry Transformed Normal Sorrow into Depressive Disorder*, New York: Oxford University Press.

Kendler, K. S. and Campbell, J. (2009), Interventionist Causal Models in Psychiatry: Repositioning the Mind-Body Problem, *Psychological Medicine*, 39:881–887.

Levin, J. (2016), Functionalism, *The Stanford Encyclopedia of Philosophy*. E. N. Zalta, ed. URL= https://plato.stanford.edu/archives/win2016/entries/functionalism

Nesse, R. M. (2005), Evolutionary Psychiatry and Mental Health, pp. 903–927. In *The Handbook of Evolutionary Psychology*. D. Buss, ed. Hoboken, NJ: John Wiley and Sons.

Reininghaus, U. and Morgan, C. (2014), Integrated Models in Psychiatry: The State of the Art, *Social Psychiatry and Psychiatric Epidemiology*, 49:1–2.

Ryle, G. (1949), *The Concept of Mind*, Chicago, IL: The University of Chicago Press.

Skinner, B. F. (1953), *Science and Human Behavior*, New York: The Macmillan Company.

Smart, J. J. C. (2017), The Mind/Brain Identity Theory, *The Stanford Encyclopedia of Philosophy*. E. N. Zalta, ed. URL= https://plato.stanford.edu/archives/spr2017/entries/types-tokens/

Tavakoli, H. (2009), A Closer Evaluation of Current Methods in Psychiatric Assessments: A Challenge for the Biopsychosocial Model, *Psychiatry*, 6:25–30.

Wetzel, L. (2014), Types and Tokens, *The Stanford Encyclopedia of Philosophy*. E. N. Zalta, ed. URL= https://plato.stanford.edu/archives/spr2014/entries/types-tokens/

4

"Aren't Mental Disorders Just Chemical Imbalances?," "Aren't Mental Disorders Just Brain Dysfunctions?," and Other Frequently Asked Questions about Mental Disorders

Ginger A. Hoffman

During my introductory course in philosophy of psychiatry, many of my students invariably pose the questions: "Aren't mental disorders just chemical imbalances?" and "Aren't mental disorders just brain dysfunctions?" These questions are usually offered as responses to our overarching discussions about the nature of mental disorders—for example, as affirmative responses to the specific question of whether mental disorders are simply bona fide illnesses "just like diabetes." Such responses are sensible and inevitable, given their coherence with our society's current dominant characterization of mental disorders (also referred to as "mental differences," "mental illnesses," "psychiatric disabilities," and "psychiatric conditions").[1]

Consider the following (representative) personal account:

> When I was 17 years old, I got so depressed that what felt like an enormous black hole appeared in my chest. Everywhere I went, the black hole went too. So to address the black-hole issue, my parents took me to a psychiatrist at Johns Hopkins Hospital. She did an evaluation and then told me this story: "The problem with you," she explained, "is that you have a chemical imbalance. It's biological, just like diabetes, but it's in your brain. This chemical in your brain called serotonin is too, too low. There's not enough of it, and that's what's causing the chemical imbalance. We need to give you medication to correct that." Then she handed my mother a prescription for Prozac. (Spiegel 2012)[2]

This chapter's aim is to answer the two title questions (henceforth considered together as "the brain/chemical question") and to explore the resulting answers'

implications for associated questions, concerns, and assumptions—in particular, for the larger question of whether psychiatric conditions are bona fide illnesses "just like diabetes." To preview its conclusions, it answers the brain/chemical question with agnosticism, but further maintains that *even if* psychiatric conditions are, in fact, robustly associated with chemical and brain differences, this alone would not necessarily entail that psychiatric conditions are bona fide *illnesses* or *disorders.* As such, this chapter reveals that the brain/chemical question, whether its questioners realize it or not, is a fundamentally philosophical one, and that its answer therefore must rely heavily on work in philosophy of psychiatry and philosophy of mind—not just neuroscience.[3]

As one approach to answering the brain/chemical and accompanying questions, this chapter canvasses what we can and can't conclude about psychiatric conditions given the presumed presence of brain or chemical differences associated with them. Thus, another way to phrase the main question of this chapter is: "What can the results of neuroscientific and neurochemical studies *really* tell us about psychiatric conditions?"

Relevance of the brain/chemical question to the philosophy of psychiatry

How is answering the brain/chemical question important for philosophy of psychiatry, and important in general? First, note that one of the central questions within philosophy of psychiatry is the aforementioned question of what psychiatric conditions *are*: Are they *bona fide* illnesses "just like diabetes," or are they non-pathological forms of neurodiversity worthy of respect? (Or something in-between? Or something else entirely?). Depending upon how the brain/chemical question is interpreted, a positive answer to it *may* support the former illness option. Thus, whether such a positive answer is true, and whether it *actually* provides such support, has clear consequences for this central debate in the philosophy of psychiatry. And, since philosophy of psychiatry is a discipline intertwined with our day-to-day lives, it also carries deeply important personal and political ramifications. For example, thinking of oneself as ill versus divergent can alter one's self-conception, self-respect, self-acceptance, and prospects for leading a flourishing life. And the way society categorizes people with so-called psychiatric conditions (e.g., as ill versus divergent) has clear implications for the stigma they encounter, the funding the state is willing to provide for their treatments, accommodations, and/or disability support, and the views others

hold about their levels of competence, responsibility, and blameworthiness, as just a few examples.

Another way in which the brain/chemical question intersects with philosophy of psychiatry is via its prima facie challenge to the relevance of the discipline itself. That is, if psychiatric conditions are "just" chemical imbalances/brain dysfunctions, then one might wonder why we even need philosophy of psychiatry to begin with: Wouldn't the pressing puzzles about psychiatric conditions best be solved by scientists? There are two immediate answers to this. First, even *if* all aspects of psychiatric conditions were best directly investigated within the realm of science, this doesn't mean that the philosophical investigation of science itself would be obviated. Consider the importance of both philosophy of science and bioethics to "purely" scientific questions within fields like physics, genetics, and biomedicine. Second, the brain/chemical question fails to vanquish philosophy of psychiatry because it itself is a philosophical question. For instance, it demands answers about the relationships between the mind and the brain—a central venture of philosophy of psychiatry and philosophy of mind more generally.

In what follows, therefore, different variations of the brain/chemical question will sequentially be examined, along with several implicit questions and assumptions that accompany them.

Starting question: Don't we *already know* that psychiatric conditions are just chemical imbalances and/or brain dysfunctions?

Often, the brain/chemical question is phrased in such a way as to imply that "we" (scientists and, by extension, the general public) *already* have an affirmative answer. This framing provides a useful starting point for our following exploration of the panoply of interpretations of this question. However, we will start by setting aside two elements of it:

1. The phrase "are just," which seems to imply some sort of reductionism—perhaps insinuating that everything about psychiatric conditions can be comprehended *fully* in scientific terms. We will return to address this possibility in our discussion of Question 3.
2. The notion of an *im*-balance or *dys*-function. These prefixes carry an important normative assumption (i.e., that the chemical and brain differences are *bad*) that deserves examination in its own right—this is the

task of answering Question 4. Until then, we will consider the chemical and brain *differences* that may be associated with psychiatric conditions, shedding any assumptions about the goodness or badness of such differences.[4]

Setting aside "are just" and "im/dys" yields Question 1.

1 Question 1 (Starting question, modified): Don't we already know that psychiatric conditions are consistently associated with certain chemical and/or brain differences?

Let's first focus on the issue of chemical differences, and then on the issue of brain differences.

1.1 Don't we already know that psychiatric conditions are consistently associated with certain chemical differences?

The short answer is "not really." This isn't to say that there is *no* evidence for average chemical differences between people with certain psychiatric conditions and people without, but this evidence (I think most psychiatrists and neuroscientists would readily agree) is far from conclusive. This is perhaps surprising, given the ubiquity of the notion that psychiatric conditions are chemical imbalances. But first, note that, when someone is diagnosed with depression, or schizophrenia, or attention deficit disorder, etc., their diagnosis is not based on any measurement of their individual "chemical" levels—there is no blood or other laboratory test performed. Valenstein makes this point when dismantling the common analogy that psychiatric "drugs treat mental disorders in the same way that insulin treats diabetes" (Valenstein 1998: 221). He writes:

> When insulin is prescribed for a diabetic it is only after a reliable test has measured the extent of the patient's glucose metabolism problem, from which, in most instances, an insulin deficiency can be inferred. The dose of insulin that is prescribed can be based on a reasonably good estimate of the magnitude of the deficiency. Moreover, we also have a good understanding of how insulin regulates glucose metabolism and how a deficiency of that hormone can produce diabetic symptoms. In sharp contrast, a psychiatrist performs no laboratory test to determine if a mental patient has any chemical deficiency or excess. (Valenstein 1998: 222)

What's more, a psychiatrist wouldn't even *know* for certain what type of laboratory test to perform, since there is no current consensus regarding which chemical differences would reliably signify different psychiatric conditions.

Again, however, this is not to say that there isn't *any* evidence for chemical differences associated with psychiatric conditions. For example, in the case of depression, there have been a number of studies indicating differences in the levels of neurotransmitter metabolites (specifically, the products of broken-down neurotransmitters) in the blood or cerebrospinal fluid of depressed persons versus that of neurotypicals. But: this is quite an indirect measure of what is happening in the brain, and is an even more indirect measure of what may be happening in certain brain regions. Neuroscientists and psychiatrists have long noted this indirectness, so much so that it's arguably safe to conclude that a simplistic "chemical imbalance" theory[5] is outright rejected by current wisdom in neuroscience and psychiatry. Here are two representative quotations:

1. According to Ronald Pies, a psychiatrist: "In truth, the 'chemical imbalance' notion was always a kind of urban legend—never a theory seriously propounded by well-informed psychiatrists" (2011).
2. From what is arguably the most authoritative and popular reference on psychopharmacology used by psychiatrists: "There is no clear and convincing evidence that monoamine [a type of neurotransmitter or "chemical"] deficiency accounts for depression—i.e., there is no 'real' monoamine deficit" (Stahl 2013: 267).[6]

However, one might still insist that we have ample evidence of a chemical difference from a different source: namely, from the fact that many successful drug treatments for psychiatric conditions work by altering one's "chemical balance."[7] But, just because a drug *causes* a change in one's chemical balance doesn't mean that there was a chemical *im*balance or difference present to begin with. Johanna Moncrieff, a psychiatrist, offers a terse example: "Alcohol can be helpful for social anxiety, but that's not because people have an alcohol deficiency" (Moncrieff 2016: 157).[8]

1.2 Don't we already know that psychiatric conditions are consistently associated with certain brain differences?

The answer depends upon how stringent we are in our definition of "consistent."[9] If we are willing to define "consistent" loosely,[10] the answer here may indeed be "yes." At least, the evidence that neurodivergents (on average) have different brain features than do neurotypicals (on average) is arguably more convincing than that for chemical differences. Many would contend that an enormous number of studies (over 15,000 [Krystal and State 2014: 206]) have indeed yielded some general trends in brain differences between neurodivergents and neurotypicals.[11]

For example, again focusing on the case of depression, three brain regions seem to be quite commonly altered in diagnosed persons compared to neurotypicals: the amygdala, the hippocampus, and the frontal cortex (including many sub-regions like the anterior cingulate cortex and the dorsolateral prefrontal cortex) (Higgins and George 2013: 252–253).[12]

However, several caveats exist. Here are two, starting with the more subtle and following with the more striking. First, the differences reported in these studies are almost always differences between the *average* brain feature of a group of neurodivergents and the *average* brain feature of a group of neurotypicals. Imagine a study looking for differences in levels of activity in Brain Area X in two groups of people: persons diagnosed with depression and neurotypicals. Each group may be composed of individuals who all manifest very different levels of activity of Brain Area X, such that many of these individuals will fall far from the average for their group. In fact, it is even possible that many neurotypical individuals exhibit Brain Area X activity levels characteristic of the "average depressed brain," and that many diagnosed individuals exhibit Brain Area X activity levels characteristic of the "average neurotypical brain."

For example, Simons (2018) reports exactly this sort of result in a recent study of ADHD (Hoogman et al. 2017). This study reported that the average volumes of several brain areas in a group of individuals diagnosed with ADHD were statistically smaller than the average volumes of the corresponding brain areas in a group of neurotypicals. Despite the statistical significance between the *averages* of these two groups, each group was composed of individuals with a very wide array of volumes of the relevant areas, such that any given individual in either group was likely to possess a volume that was shared by an individual in the other group. In other words, there was substantial overlap of the ranges of the relevant brain areas' volumes between the two groups. Simon writes: "Only a very few outliers had brain volumes that were actually outside the variability of the two groups. Thus, out of 100 people with ADHD, only five could be expected to have brain volumes that were not equal to those of people in the 'healthy' control group" (Simons 2018; see also Holmes and Patrick 2018).

The second, more glaring, caveat is that not all of these studies yield the same results; as mentioned above, their consistency is "loose," at best. Consider, again, the example of depression. In the neuroscience and psychiatry textbook mentioned above, Higgins and George conclude, "Still, there is no consensus regarding the CNS [central nervous system] alterations in the depressed brain" (2013: 253). This lack of consensus is emphatically underscored by a recent study of ninety-nine neuroimaging experiments published from 1997 to 2015

that, to the authors' knowledge, represents "the largest metaanalysis of task-based neuroimaging experiments in UD [depression] to date" (Müller et al. 2017: 53). Across these ninety-nine studies, there were *no* reliable average brain differences between persons diagnosed with depression and neurotypicals that reached the level of statistical significance.[13, 14] And there is reason to think that such inconsistencies extend beyond depression to other psychiatric disorders (see Müller et al. 2017: 54; Weinberger and Radulescu 2016).

Moreover, several studies question the dependability of one of the main methodologies (MRI) used to generate much of the brain difference data (Eklunda et al. 2016; Weinberger and Radulescu 2016). Weinberger and Radulescu go as far as saying:

> We opine that current studies are plagued by so many possible systematic confounders that one can only wonder whether, like Wolfgang Pauli, "these results are not only not right, they are not even wrong!" We would caution that researchers and clinicians pause and rethink carefully the conclusions that can be drawn from these various MRI findings in psychiatric research. (2016: 32)[15]

CONCLUSION FROM CONSIDERATION OF QUESTION 1:

WE MAY OR MAY NOT "ALREADY KNOW" THAT SOME PSYCHIATRIC CONDITIONS ARE CONSISTENTLY ASSOCIATED WITH SOME BRAIN DIFFERENCES; THIS WILL DEPEND IN PART ON HOW STRINGENT OUR STANDARDS FOR "CONSISTENCY" ARE. NOTABLY, THERE IS EVIDENCE CHALLENGING THE COMPLETE CONSISTENCY OF THE RELEVANT FINDINGS. THE CONSISTENCY OF THE ASSOCIATIONS BETWEEN PSYCHIATRIC CONDITIONS AND CHEMICAL DIFFERENCES IS ARGUABLY EVEN LESS ESTABLISHED.

However, it's certainly possible that continued research will give us a clearer, more unified, and more consistent picture of brain and chemical differences between neurodivergents and neurotypicals. This leads to Question 2.

2 Question 2: Won't we be able to know (in the future, once science advances) that psychiatric conditions are consistently associated with certain chemical and/or brain differences?

The answer here may very well be "yes." It may be that, once science advances, the lack of consistency mentioned above will be resolved and we will be able to confidently say that particular psychiatric conditions are consistently associated

with one or more particular brain features—for example, that obsessive compulsive disorder (OCD) is consistently associated with increased activity in the basal ganglia. On the other hand, there may be reasons that this will never be the case, perhaps owing to things like, in this example, the complexity of OCD, the vast individual variations in brain functioning, and what may turn out to be insurmountable limitations of scientific measurement. Nevertheless, it is not unreasonable to think that we will be able to eventually confidently assign one or more specific brain features to certain disorders. *If* this is the case, we will, indeed, be able to know that psychiatric conditions are consistently *associated with* chemical and/or brain differences. But: does this mean that they *are* chemical and/or brain differences? And, more importantly, what else, exactly, would we be warranted to conclude from this?

So, perhaps surprisingly, Questions 1 and 2 are not really the important questions—at least, not in terms of what psychiatric conditions *are* and what they *mean*.[16] For example, even if a consistent, reliable association is discovered in the future, this association *on its own* doesn't tell us very much, if anything, about what psychiatric conditions are, nor about what it means to be diagnosed with one. Nor does it determine whether psychiatric conditions are *bona fide* illnesses (this will be explored more under Question 4). We now turn to examine these more vital questions, starting with the question of whether psychiatric conditions *are just* certain brain differences (Question 3).

CONCLUSION FROM CONSIDERATION OF QUESTION 2:

ONCE SCIENCE ADVANCES, WE MAY BE ABLE TO KNOW THAT PSYCHIATRIC CONDITIONS ARE CONSISTENTLY ASSOCIATED WITH CERTAIN CHEMICAL AND/OR BRAIN DIFFERENCES, BUT, IMPORTANTLY, CRUCIAL QUESTIONS WOULD REMAIN ABOUT WHAT SUCH ASSOCIATIONS MEAN.

3 Question 3: Won't we be able to know (in the future, once science advances) that psychiatric conditions *are just* certain brain differences?

Two things have happened in the transition from Question 2 to Question 3. First, the phrase "are just" has reappeared, replacing the phrase "are associated with": it is now time to directly address the issue of what psychiatric conditions *are.* Second, the question of *chemical* differences has been set aside for reasons of space.[17]

3.1 Review of possible types of mind-body relationships

Answering Question 3 requires an understanding of how the mind and body (including the brain) are related. And this understanding requires a familiarity with canonical work in the philosophy of mind. To see in more detail why such a familiarity is necessary, first, note that psychiatric conditions consist of various mental states, processes, and other features (often described as symptoms), including mental states like emotions, moods, beliefs, memories, desires, and mental processes like attention and cognition. And note that the brain differences commonly associated with psychiatric conditions consist of various brain states (and other brain features) (see endnote 4). Thus, in order to ascertain whether psychiatric conditions "are just" brain differences, we must ascertain whether mental states "are just" (or, are *identical to*) brain states. And this is a fundamentally philosophical question. In other words, philosophy is the discipline directly responsible for offering methodical, logical arguments for and against the possibility of different types of relationships between the mind and the brain. So here is a very preliminary sketch of the two major philosophical positions concerning the relationship between the mind and brain, the latter of which is divided into two subspecies.

Dualism: Mental states and brain states are made up of two different types of substances: immaterial and material, respectively. Thus, they cannot be identical to one another.

For example, mental states may be "soul-like" or composed of another type of non-physical "stuff," separate from atoms and other physical matter that make up the brain. However, according to most forms of dualism, even though mental states and brain states are composed of two distinct types of metaphysical ingredients, they can and do causally interact with one another. That is, the brain affects the mind, and vice versa. For example, changes in the brain (e.g., changes in neurotransmitter levels caused by antipsychotics) can cause changes in the mind (e.g., a dampening of hallucinations) and changes in the mind (e.g., thinking of something for which you are grateful) can cause changes in the brain (e.g., again, changes in neurotransmitter levels in a specific brain region).

Physicalism: Mental states are identical to brain states, in much the same way that water is identical to H_2O.

Physicalism can be divided into (at least) two subspecies: type physicalism and token physicalism. If unspecified, "physicalism" usually refers to type physicalism. To understand the difference between these

two subspecies, first consider the (admittedly rather abstract) distinction between types and tokens. A token is an individual instance or occurrence of something, whereas a type is a class whose members are tokens that are all exactly alike one another. This distinction is much clearer with an example. Take the word "poodle." This word has *five* types of letters (letters of types "p," "o," "d," "l," and "e"), but six *tokens* of letters (there are two tokens of the type of letter "o," and one token each of the other four types of letters).

Applying this to the two subspecies of physicalism, type physicalism states that " … mental types are identical to physical types" (Mandik 2014: 133) and token physicalism states that " … every mental token is identical to a physical token." (Mandik 2014: 133). In slightly more detail:

Type physicalism: A type of mental state M is identical to a type of brain state B. This means that every instance (token) of M will always be mental type M *and* brain type B.[18] For example, type physicalism might say that the type of mental state "pain" is identical to activity in the periaqueductal gray region of the brain, and that this would be true in all individuals, in all contexts.

Token physicalism: "Every mental token is identical to a physical token" (Mandik 2014: 133). So, one token of mental state type M may be identical to a token of one type of brain state (type B) in one context, while another token of mental state M may be identical to a token of a *different* type of brain state (type C) in another context. For example, token physicalism might say that a token of the type of mental state pain in Heidi is identical to low activity in Heidi's brain region P, while another token of the *same type of* pain in Lavelle is identical to moderate activity in Lavelle's brain region Q.[19]

3.2 Returning to Question 3: Won't we be able to know (in the future, once science advances) that psychiatric conditions *are just* certain brain differences?

Having introduced some basic positions in the philosophy of mind, we can now see that the answer to Question 3 depends partly upon which of these philosophical positions is correct. That is, the word "are," in the context of Question 3, can be taken to mean "identical to," so the answer to Question 3 will depend at least partly upon whether physicalism is true. Unfortunately or fortunately, however, there is currently no definitive "winning position" in the debate over the nature of the mind-brain relationship. In terms of *popularity*,

many (but clearly not all) psychiatrists and philosophers today seem to favor some type of physicalism, because they are skeptical of the existence of a "soul-like" or non-physical substance. But, popularity doesn't always imply truth, and strong arguments supporting any of the above positions can be found. Thus, it is best to explore possible answers to Question 3 sequentially in the context of each of the philosophical positions presented above.

First, if dualism is true, then we clearly cannot conclude that mental differences composing psychiatric conditions "are just" brain differences. This is because, according to dualism, mental states are metaphysically distinct entities from brain states. For example, the presence of a type of intrusive thought in a psychiatric condition might be *associated with* or *causally related to* a certain brain difference, but it wouldn't be accurate to say it is "identical to" or "just" a brain difference.

If physicalism is true, however, then we are in a better position to possibly conclude that psychiatric conditions are, indeed, (just) brain differences. However, whether such a conclusion is warranted will depend largely upon whether type or token physicalism is true. Consider the following hypothetical example of a mental difference: the presence of a feeling of hopelessness in several persons diagnosed with depression, and its absence in several neurotypicals. Now, consider **type physicalism.** Assuming that hopelessness is a type of mental state, if type physicalism is true, then any instance (token) of hopelessness will be identical to an instance (token) of a distinctive type of brain feature—call it "Brain State H." In this case, it would arguably make sense to say that the mental state of hopelessness *is* Brain State H. And, because the same sort of story would presumably be true for other mental states that characterize depression, then it would arguably make sense to say that depression itself *is* a collection of certain brain features—one could call this collection "Brain Pattern D."

So, if type physicalism is true, then it would arguably make sense to say that psychiatric conditions are brain differences and, moreover, are certain distinctive types of brain differences. But does it make sense to go further and say that they are *just* (certain types of) brain differences? Clearly, it depends on what is meant by "just." "Just" may imply, for example, that how the mental/physical state in question (e.g., hopelessness) *feels* to the individual experiencing it is unimportant. This extreme possibility, however, does not follow from most interpretations of type physicalism. On the contrary, it is certainly possible that mental and physical states are identical, just like water is identical to H_2O, and that the experience of them *as* mental nevertheless remains important.[20] Thus, because the word "just" may open itself up to such unwarranted and overly

reductionistic interpretations, it is safest to say that, if type physicalism is true, then psychiatric conditions are brain differences (dropping the "just").

Turning to **token physicalism,** we can see that not all physicalists need to accept that the mental states making up certain psychiatric conditions are necessarily identical to certain *types of* brain differences. To see this, reconsider the hopelessness example. If token physicalism is true, then an instance of hopelessness in one person at one time could be identical to, say, Brain State H, while another instance of hopelessness in a different person, and/or at a different time, could be identical to, say, Brain State I. Moreover, if token physicalism also applies to the other mental states that characterize depression, then the manifestation of depression in one person could be identical to a very different collection of brain features (call it "Brain Pattern D") than the manifestation of the same type of depression in another person (call it "Brain Pattern E"). In other words, if token physicalism were true, depression could be identical to *different* sorts of brain differences in different individuals and different contexts, and it would therefore be a stretch to say that it (and other psychiatric conditions) were certain types of brain differences.

Finally, token physicalism, if true, could provide a counterintuitive twist on what we tend to automatically conclude from the idea that there are brain differences between neurodivergents and neurotypicals. Consider the following. Many of us assume that, if neuroscientists found a difference (between two individuals) in activity levels in a particular area of the brain often associated with anxiety (call it "Brain Area A"), then this would mean that one individual (call him Randolph) with increased activity in Brain Area A would presumably have pronounced anxiety compared to the other individual (call her Rashmi) who had typical activity in Brain Area A. But this conclusion may not be warranted (for many reasons). One reason is that it is actually possible, if token physicalism and certain other conditions are true, that a brain difference such as this one could actually provide evidence of a mental *similarity* between these two individuals. For example, it could be that a certain type of anxiety in Randolph is identical to heightened[21] activity in Brain Area A, while the *same type* of anxiety in Rashmi is identical to heightened activity in a different brain area—Brain Area X (and is not associated at all with Brain Area A). In this case, witnessing heightened activity in Brain Area A (along with typical activity in Brain Area X) in Randolph, and typical activity in Brain Area A (along with heightened activity in Brain Area X) in Rashmi, would *fail to* indicate a difference in their relative anxiety levels, and would instead be consistent with a mental *similarity*. In fact, we would expect to see exactly this result when one and the same type of anxiety

is occurring in both Randolph and Rashmi. And the same could be argued for the hypothetical examples of hopelessness and Brain Areas H/I above, as well as for a host of other brain and mental states and psychiatric conditions.[22]

This possibility is extremely important to recognize. Usually, brain differences are unreflectively taken to showcase *deep and substantive* evidence of mental differences between neurodivergents and neurotypicals. Indeed, they are often (confusingly) assumed to indicate not only that mental differences between these two groups *exist*, but that they are in fact *real* and *legitimate*. Here is one real-life example: some individuals have sought the counsel of Dr. Amen—a neuroimaging expert—to tell them if they are "really" depressed. He offers to provide this information by imaging their brains and seeing whether they exhibit the "characteristic depressed brain difference" (Amen 1998). There are many deeply flawed aspects of this enterprise; one is the failure to recognize that, as just explained, the presence of a brain difference (e.g., a difference in Brain Area A, or a difference in Brain Area X) might not indicate the presence of a mental difference *at all*.[23] So Dr. Amen's enterprise, along with other popular assumptions about the power of neuroscience to demonstrate a mental difference's "reality," are intellectually (and perhaps ethically) disturbing for this and other reasons. Importantly, as shown above, a basic understanding of positions in the philosophy of mind could avoid and dissolve this sort of disturbance.[24]

CONCLUSION FROM CONSIDERATION OF QUESTION 3:

THE ANSWER TO QUESTION 3 WILL DEPEND UPON WHAT VERSION OF THE MIND-BRAIN RELATIONSHIP IS TRUE. HOWEVER, *NO* VERSION STRICTLY IMPLIES THAT PSYCHIATRIC CONDITIONS ARE *JUST* CERTAIN BRAIN DIFFERENCES, IF "JUST" MEANS THAT MENTAL FEATURES AND INDIVIDUAL EXPERIENCES ARE UNIMPORTANT.

While Question 3 laid out the possible metaphysical status of brain differences, Questions 4 to 7 turn more toward investigating the broader implications and assumptions of this status; these questions focus on what we can or can't conclude from the presence of brain differences. In considering them, it's important to keep in mind the possibility raised directly above—that, in certain situations, if token physicalism is true, then certain brain differences may not even correspond to mental differences in the first place. However, it is also possible that other certain brain differences *do* indicate mental ones. Thus, we will examine Questions 4 to 7, for the sake of argument, under the assumption that the brain differences under consideration do, in fact, correspond to the mental differences in question.

4 Question 4. Don't the brain differences underlying[25] mental differences (that compose psychiatric conditions) show that these mental differences are indeed *dysfunctional, diseased, or disordered*—i.e., that there is something *wrong* with such mental differences?

Here, the second issue bracketed in Question 1 is re-introduced: many people assume that, if a brain difference can be shown to "underlie" a mental difference, then this means that this mental difference is indeed *dysfunctional, diseased*, or *disordered*. In other words, brain (or chemical) differences are often taken to be clear evidence of mental *pathology*. And this is largely because the relevant brain (or chemical) differences are assumed to be signs of brain *dys*function to begin with.

It is best to tackle this question in two steps. First, Section 4.1 will argue that a brain difference does not, *on its own*, automatically indicate mental pathology. Next, Section 4.2 acknowledges that it is possible, though, that a brain difference that does happen to be a brain dysfunction may indeed indicate a mental dysfunction. However, Section 4.2 will further argue that we don't know enough to establish that most brain differences associated with psychiatric conditions are actually brain dysfunctions in the first place. Moreover, there are notable hurdles to attaining such knowledge in the future.

4.1. Don't brain differences, on their own, indicate mental pathology?

Here, we are considering cases where it is unclear whether the brain differences in question are actually brain dysfunctions. It's fairly straightforward to see how, in these cases, such brain differences by themselves do not necessarily indicate any corresponding mental pathology. Of course, brain differences are certainly *consistent* with mental pathology, but their existence clearly does not necessitate it. This is because, simply put, all or most mental phenomena will almost certainly have some brain differences underlying them. And it's certainly not the case that all or most mental phenomena are pathological! When we think a thought, this will presumably be associated with/identical to a brain difference; when we hear a sound, or see something, or feel fascinated, or bemused, or devastated, or remember to feed the guinea pigs, or think or feel or perceive or remember *anything*, presumably it will be associated with/identical to some brain difference or other.

To spell out this point more thoroughly in reference to basic positions within the philosophy of mind, we can see, first, if physicalism (type or token) is

true, then every mental state—dysfunctional and functional, pathological and healthy—will be identical to some brain state or other. And if dualism is true (assuming favored forms of dualism that allow for causal interactions between the mind and the brain, or even for a non-causal *correlation* between the mind and the brain), then all mental states will almost certainly be at least *associated* with some type of brain activity or state.

The only way this wouldn't be the case would be if some unpopular forms of dualism (which deny interactions and/or other forms of correspondence between the mind and brain) were true. But, these forms are unpopular precisely because we have abundant evidence that the mind and brain *do*, at least, interact (or, are identical to one another). If this isn't already obvious, consider the following two examples. First, changes in the mind (e.g., remembering a sad song) tightly correlate with, and thus almost certainly cause (or are) changes in the brain (e.g., they may be changes that cause tears to come to our eyes. If the brain didn't change as a result of this sad memory, it would be hard to account for the physical result of crying). And certain changes in the brain, likewise, lend evidence to the idea that the brain is *not* completely divorced from the mind [e.g., directly changing one's brain with LSD *at least* tightly correlates with (or effects, or even constitutes) changes in one's mental activity]. Thus, all sorts of mental differences—healthy or not—will correspond in some way to brain differences, on all plausible models of the relationship between the mind and the brain. Therefore, the existence of some brain difference or other, *as such,* does nothing to indicate a dysfunction.

4.2. Can't we tell that certain brain differences are brain dysfunctions? And can't we then show that the corresponding mental differences are indeed dysfunctional, diseased, or disordered?

One way a brain difference could provide evidence of a mental dysfunction could start with the fact that this brain difference itself was dysfunctional. But in order to establish whether a brain difference is itself dysfunctional, much more would have to be known about what counts as "functional" brain activity to begin with. Currently, however, knowledge in this area is scant. Moreover, there are significant logistical and conceptual hurdles to obtaining such knowledge in the future. To see this, imagine how we might determine a "functional" range of activity of a particular brain area. Say that Brain Area A has been observed to display anywhere from zero to one hundred "units" of activity.[26] How would we

go about determining the boundaries of *functional* activity? Maybe it should be 25 to 75? Or anything that is within one standard deviation of the mean? Or two standard deviations? Any non-arbitrary answer, it seems, must depend upon the range of "functionality" (or "normality") of the associated *mental* state. For example, if activity in Brain Area A is consistently associated with anxiety, and an activity level of 95 usually corresponds to what is considered "high anxiety" or "panic," then perhaps we can reasonably conclude that an activity level of 95 in Brain Area X is dysfunctional brain activity.

However, there are two things to note about this. First, Question 4 is presumably asking whether neuroscience can offer *additional* evidence of pathology—i.e., evidence over and above what is already known about the psychological and mental aspects of the psychiatric condition in question. But in the above scenario, neuroscientific data is not providing *additional* evidence. Rather, any evidence it is supplying is derivative from the mental evidence already known. In the above scenario, evidence for brain dysfunction is inferred from presumed mental dysfunction; thus, any attempt to then derive more evidence for mental dysfunction from that very brain dysfunction would clearly be circular.

Second, the judgment about mental functioning in the above strategy may itself be shaky insofar as it is likely to be deeply infused with values and social norms. For example, how does one know how much anxiety is "too much"? Is it when the individual is suffering? If so, this will depend upon the individual. And what level of suffering (if any) counts as dysfunctional? Is it when there is *any* anxiety? Clearly, though, some anxiety seems "functional" in our world, depending on the context. High anxiety during an important test, or when being chased by a hyena, seems not only acceptable, but, in fact, seems to be a sign of *functioning*.[27] This point has been made by Jerome Wakefield when discussing activation of brain circuits:

> Particularly pernicious is the lazy notion that disorder is simply high circuit activation. Anyone who has been terrified at imminent danger or experienced an orgasm knows that this can't be right. One might object that … atypical or impairing high activation [is] disordered. But, depending on how you select your dimensions, you can make anything atypical. It is statistically typical to sleep, but the circuit activation during sleep is highly deviant from normative circuit status when awake. (Wakefield 2014: 39)[28]

CONCLUSION FROM CONSIDERATION OF QUESTION 4:

THE EXISTENCE OF A BRAIN DIFFERENCE, IN ITSELF, DOESN'T IMPLY THAT SOMETHING IS "WRONG" WITH THE CORRESPONDING

BRAIN OR MENTAL STATES, NOR DOES IT IMPLY THE PRESENCE OF A BRAIN OR MENTAL *DYSFUNCTION, DISORDER*, OR *IMBALANCE*. THE EXISTENCE OF A BRAIN *DYSFUNCTION*, ON THE OTHER HAND, MIGHT INDICATE THE PRESENCE OF A MENTAL *DYSFUNCTION, DISORDER*, OR *IMBALANCE*, BUT THERE ARE SIGNIFICANT CONCEPTUAL HURDLES TO ASCERTAINING WHICH BRAIN (AND MENTAL) DIFFERENCES ARE ACTUALLY DYSFUNCTIONS IN THE FIRST PLACE.

5 Question 5. Don't the brain differences that neuroscience has revealed (and will continue to reveal) show that the associated psychiatric conditions are initially *caused* by something "biological"—e.g., genetic variations, brain lesions?[29]

Before directly answering this question, it's critical to first note that we already have substantial scientific evidence that virtually[30] no psychiatric conditions are initially (exclusively) caused by something biological (Higgins and George 2013: 59–60; Plomin et al. 2016; Polderman et al. 2015).[31] So technically we already know that brain differences cannot "show" this, because it's not true in the first place!

Second, however, in order to more directly and fully answer this question, it's important to try to unpack what it's really asking. Although different people ask this question for different reasons, my suspicion is that many of them hope that unveiling a biological cause will lend credence to the idea that the mental differences in question are, indeed, diseases (see Question 4). This makes sense: when we think about diseases in general, we note that they are often *defined* by their biological causes. Huntington's disease is defined by appeal to the genetic mutation that causes it, strep throat is defined by appeal to the *Streptococcus pyogenes* bacterium that causes it, etc. So even if the presence of a brain difference *as such* doesn't imply that the mental difference in question is a disease or dysfunction (as discussed above under Question 4), then (one might think) perhaps the presence of a biological *cause* would affirm its disease status.

But here, again, we can illustrate that (1) the brain differences that neuroscience has revealed do not by themselves necessarily show that those differences' initial causes were biological. In fact, neither differences in brain function nor structure necessarily imply the presence of a genetic or biological initial cause. And, more importantly, (2) even if those brain differences' causes *were* biological, this would not entail that the mental differences with which they correspond were indeed diseases.

First, take (2), and focus on the idea of a genetic cause. Even if a particular effect has a 100 percent genetic cause (i.e., even if it is fully determined by variations in a particular gene or genes), this doesn't mean that it is a disease. Consider eye color. This is (more or less) 100 percent determined by variations in a particular gene, but it's clear that having brown eyes, or blue eyes, or green eyes, etc. is not a disease.

Now consider (1). As stated above, as long as we assume the truth of physicalism or most forms of dualism, differences in brain function will correspond to *all sorts of mental states*, many of which do *not* have primarily biological causes. Clearly, many of the thoughts, feelings, sounds, sights, and memories we experience will not be solely caused by biological factors like genes or viruses. Instead, their causes will often be changes in the environment. Consider a simple example: when I look at a gray balloon, and then shift my gaze to look at a circus tent, the activity in part of my brain (my visual cortex) changes. (Hoffman 2012: 34). Here, aspects of the environment (the balloon; the tent) are important causes of changes in brain function.[32]

This makes sense because the brain, of course, is configured to accept myriad types of inputs from the environment. Not only do the five sense organs have more or less direct connections with the brain, but the brain is receiving a continuous flow of blood that may usher in things like increased levels of "stress hormones," themselves determined by levels of stress in one's environment (Sapolsky 1996). All such inputs can and do serve as significant causes of brain differences.

Furthermore, "non-biological" causes like the environment can change not only brain *function,* but brain *structure*. Thus, any given feature of brain structure that we see associated with psychiatric conditions may at least theoretically be the result, or the partial result, of the environment. As Hoffman (2012) writes:

> There is an abundance of empirical evidence showing that the environment can have profound effects on brain structure. The environment has been shown to alter the following: the overall weight and size of the brain, neurogenesis (the "birth" of new neurons), gliogenesis (the "birth" of new glial cells), dendritic branching (the number of branches one part of a neuron has—thus, the shape of a neuron), synaptogenesis [the formation of new synapses …] (Mora et al. 2007: 80–81. See also: Pascual-Leone et al. 2005; Ming and Song 2005, and Jacobs et al. 2000). (Hoffman 2012: 34–35)

As just one example, there is evidence that "enriched" environments for rats (with more space, toys, ramps, etc.) cause increases in all of the aforementioned structural features (Mora et al. 2007: 80–81). Inversely, certain types of stressful environments have been correlated with *reduced* neurogenesis and other

structural changes. For example, Sheline remarks: "There is also evidence that stress produces neuronal damage and structural changes in animal models and potentially in humans, although to date evidence in humans derives from studies of posttraumatic stress disorder" (2003: 346; see also Hoffman 2012: 34–35; Jacobs et al. 2000; Levin et al. 2007: 223). And several other canonical studies attest to the impact the environment may have on brain structure: as just two more examples, there is good evidence that gray matter volume in humans is altered by learning to juggle (Draganski et al. 2004) and structural changes in the hippocampi of taxi drivers are associated with the acquisition of navigational knowledge of taxi drivers (Maguire et al. 2000). (See also Higgins and George 2013: 95–99; Hoffman and Bluhm 2016: 721–722.)

Thus, the functional and structural brain differences that neuroscience has thus far revealed to be roughly associated with psychiatric conditions may be caused by the environment, or by genes, or (most likely) by a combination of both.[33]

CONCLUSION FROM CONSIDERATION OF QUESTION 5:

FIRST, WE ALREADY HAVE AMPLE EVIDENCE THAT PSYCHIATRIC CONDITIONS DO NOT HAVE EXCLUSIVELY BIOLOGICAL CAUSES. SECOND, THE EXISTENCE OF CONSISTENT ASSOCIATIONS BETWEEN BRAIN DIFFERENCES AND PSYCHIATRIC CONDITIONS WOULD NOT, BY ITSELF, PROVIDE SUBSTANTIAL EVIDENCE FOR GENETIC OR OTHERWISE BIOLOGICAL CAUSES OF SUCH CONDITIONS. THIRD, MOST IMPORTANTLY, A GENETIC OR OTHERWISE BIOLOGICAL CAUSE WOULD NOT NECESSARILY IMPLY THAT THE CORRESPONDING EFFECT IS A DISEASE OR DYSFUNCTION.

6 Question 6. Don't the brain differences that neuroscience has revealed (and will continue to reveal) to be associated with psychiatric conditions show that, if we want to successfully treat such conditions, then we must do so using "biological" tools like drugs?[34]

Once again, the answer is "no." Even though the mental features composing psychiatric conditions are almost certainly identical to/associated with brain features, this does not, in itself, mean that changes to those mental states (and hence their underlying brain features) must be carried out with biological tools. Ample support for this can be derived from the evidence presented in the consideration of Question 5. Because the brain is receptive to multiple types of alteration from the environment, then it is *at least* theoretically possible that

changes in the environment (e.g., the addition of various psychotherapeutic factors in one's environment) could in fact elicit changes in the relevant features of the brain (and hence the corresponding mental features). Recalling the possible philosophical positions on the relationship between the mind and the brain surveyed under Question 3, all of them (again, with the exception of some rare and implausible forms of dualism) conceptually allow that environmental factors that affect someone's mind (e.g., social supports, psychotherapy, spiritual exercises) can also affect their brain.

Here is a famous example that transcends the theoretical: many studies have provided evidence that certain forms of psychotherapy cause the same types of changes in the brains of depressed persons as do antidepressants or other drugs (Barsaglini et al. 2013; Kumari 2006; Linden 2006).[35] Although these studies may at first come across as awe-inspiring, upon reflection, it should not be at all surprising that learning how to think and feel differently via psychotherapy, along with becoming less depressed as a result of that therapy, would cause changes in the brain. Unless these parts of the mind were completely isolated from the brain (as per the rare and implausible forms of dualism mentioned earlier), then *of course* these sorts of mental changes will somehow register as brain changes. Of course, this is not to say that the possibilities are unlimited; not every type of environmental input can affect any type of brain change.[36, 37]

CONCLUSION FROM CONSIDERATION OF QUESTION 6:

THE EXISTENCE OF CONSISTENT ASSOCIATIONS BETWEEN PSYCHIATRIC CONDITIONS AND BRAIN DIFFERENCES WOULD NOT, BY ITSELF, ENTAIL THAT CHANGES TO THESE CONDITIONS COULD ONLY BE MADE "BIOLOGICALLY" THROUGH DRUGS, SURGERY, OR OTHER BIOLOGICAL MEANS.

7 Question 7. Isn't neuroscience a valuable tool for bettering the lives of those with psychiatric conditions?

Yes, for at least some individuals. This will depend upon whether individuals believe that using "tools" to change their psychiatric conditions will actually better their lives.[38] For those individuals who do wish to alleviate the palpable suffering and other problems very often associated with their psychiatric conditions, neuroscience has the potential to be extremely valuable. Even though the existence of brain and corresponding mental differences does not necessitate that such differences *must* be altered with biological means (see

Question 6), it certainly doesn't *exclude* such alteration. In fact, neuroscientific studies can allow us to discover and refine different drug therapies that may benefit individuals who find the promise of drug therapy appealing, and yet are not satisfied with current drugs (maybe, for example, because of noisome side effects). Indeed, neuroscience may enable us to develop more fine-grained tools to better help individuals achieve their own definitions of healthy, fulfilling lives, controlling exactly which aspects of their psychiatric conditions they wish to eliminate, and which they (perhaps) wish to keep.

CONCLUSION FROM CONSIDERATION OF QUESTION 7:

OF COURSE NEUROSCIENCE IS VALUABLE IN MANY WAYS. BUT PHILOSOPHY IS NECESSARY FOR FULLY UNDERSTANDING WHAT NEUROSCIENCE CAN AND CAN'T TELL US ABOUT PSYCHIATRIC CONDITIONS, AND FOR GRAPPLING WITH THE MEANING OF THOSE CONDITIONS.

8. Conclusion

Neuroscience has made and is continuing to make impressive strides in advancing our knowledge of brain activity relevant to psychiatric conditions. These strides have yielded, and will continue to yield, new and better options for handling often unwanted mental states that make up these conditions. However, many people tend to assume that neuroscientific findings can definitively resolve important metaphysical questions about the nature of psychiatric conditions, along with associated ethical questions about their significance and how they should be treated. In fact, many assume that neuroscientific findings prove the truth of the mantra that psychiatric conditions are "just" chemical imbalances/brain dysfunctions, and that, as such, there is little room left over for philosophy at all.

This chapter has argued that such assumptions are wrong, and that the claim that psychiatric conditions are "just" chemical imbalances/brain dysfunctions is itself a fundamentally philosophical one. More specifically, it has argued that our current evidence of brain differences associated with psychiatric conditions fails to settle the following issues:

- Whether the associated psychiatric conditions were caused by genetic or biological factors (see Question 5)
- How the associated psychiatric conditions should be treated (see Question 6).

- *If* the associated psychiatric conditions should be treated (see Question 4). In other words, the presence of brain differences associated with psychiatric conditions does not necessarily provide additional support that these conditions are illnesses or disorders in the first place. More generally, findings from neuroscience have failed to adjudicate a central question of philosophy of psychiatry: whether so-called psychiatric conditions are illnesses/disorders ("the illness model"), and/or are forms of neurodiversity to be accepted and perhaps celebrated ("the neurodiversity model").

In fact, one of the main take-home messages of this chapter is that our knowledge about the brain and "chemical balances" is consistent with *both* an illness model *and* a neurodiversity one. Moreover, this is not simply an artifact of the relative infancy of our neuroscientific knowledge; there is good reason to believe that neuroscience will *never* officially be able to settle this question on its own (see, especially, Section 4.2).

Thus, although neuroscience is deeply important in psychiatry, philosophy is just as important, and is indispensable for fully interpreting the meaning and implications of neuroscientific results. Without philosophy, we couldn't satisfactorily challenge common assumptions about psychiatric conditions, assumptions that have profound implications for individuals' lives—both potentially liberating and potentially oppressive.[39]

Acknowledgments

I would like to thank Şerife Tekin, Robyn Bluhm, Michelle Joy, Bhob Rainey, and my excellent students in my two spring 2018 sections of PHL 286 Philosophy of Mental Illness (Saint Joseph's University) for their helpful and insightful comments and suggestions.

Notes

1 The terminology chosen to describe mental differences is highly contentious. By "mental differences" I mean those mental and behavioral statuses that are meant to be captured and enumerated within the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (DSM), and the World Health Organization's *International Classification of Diseases* (ICD). They include

states and dispositions falling under labels including "major depressive disorder" (hereafter "depression"), "schizophrenia," "obsessive compulsive disorder" (hereafter "OCD"), "generalized anxiety disorder," "bipolar disorder," "borderline personality disorder," "autism spectrum disorder," "attention deficit hyperactivity disorder" (hereafter ADHD) etc. The reason for the controversy is that some people believe that these differences are illnesses/disorders (in other words, that they are best described with a *medical model*), some believe that they are non-pathological forms of mental and neural diversity to be accepted and celebrated (in other words, that they are best described with a *neurodiversity model*), some believe they are both, and some believe they are something else entirely. I prefer the term "mental differences" because it is compatible with all of the above positions, and is thus compatible with individuals' own preferences and beliefs. However, unfortunately, the term "mental differences" tends to be unrecognized in today's society. Because of this, my use of it may cloud the meaning of certain parts of this chapter for current readers. As such, I will use "mental differences" when the context allows, but will more commonly use "psychiatric conditions" (or, occasionally, "mental disorders") to indicate that these differences are those that are, for better or worse, the objects of psychiatric and clinical psychological practices. In employing these more traditional terms, I in no way intend to preclude the possibility of a neurodiversity model. Furthermore, when describing individuals without mental differences (so-called "sane" people, assuming such individuals even exist), I use "neurotypicals" or "typicals," since "neurotypical" is an established term within the neurodiversity community. When describing individuals with mental differences, I often use "people with psychiatric conditions," for the reasons stated above (primarily, the greater recognition value of "psychiatric conditions"). I prefer, and sometimes herein use, "neurodivergent people" or "mentally different people." However, it is worth noting at least two potential problems with this terminology. First, it is not clear that there is any defensible dividing line separating two obviously distinct groups of people: the neurodivergent and the neurotypical. Instead, it is very possible that everyone is neurodivergent in some way, or, at the very least, that there are degrees of neurodivergence that fail to admit of any clear line between "divergent" and "not divergent/typical." Second, *even if* there are two discernable categories—"the neurodivergent" and "the neurotypical"—certain studies actually provide evidence that the *majority* of the population (up to 83 percent) (Schaefer et al. 2017) meet the diagnostic criteria for psychiatric conditions at some point in their lives, and thus would *ostensibly* belong to the "neurodivergent" group (assuming that diagnostic criteria accurately capture the phenomenon of neurodiversity, which is questionable). In this case, confusingly, neurodivergent individuals would actually be statistically typical. Furthermore, even though I and others tend to use "neurodivergent" and "mentally different" interchangeably, it is neither wise, nor always accurate, to conflate the neurological

with the mental. As I discuss below, an individual may possess quite divergent/atypical brain activity, yet also be considered mentally typical. The converse is also true. However, the term "neurodivergent" has become somewhat cemented to mean "mentally divergent." Thus, this terminology certainly has its shortcomings. Nevertheless, it may currently be the *least* problematic of the available options. Finally, when more specificity is warranted (e.g., when referring specifically to individuals meeting criteria for major depressive disorder), I use both "depressed person(s)" and "person(s) with depression." There is a vexed debate about whether "person-first language" (e.g., "person with depression") is *more* or *less* respectful than "identity-first language" (e.g., "depressed person"). Arguably individuals should choose their own preferred terms for themselves. But because I am referring to individuals whose preferences are unknown (or simply to abstractions of individuals), I use both terms (more or less interchangeably) to represent both possible choices.

2 Two other representative statements include:
   1 "Depression is known to be caused by a deficit of certain neurochemicals or neurotransmitters, especially norepinephrine and serotonin." (From the self-help book titled *Change Your Brain Change Your Life* [Amen 1998: 47])
   2 "While the cause is unknown, depression may be related to an imbalance of natural chemicals between nerve cells in the brain. Prescription Zoloft works to correct this imbalance." (From a television commercial for the antidepressant Zoloft [2001]).

3 I write this as someone who is currently a philosopher, and was formerly a neuroscientist investigating basic biological mechanisms relevant to psychiatric conditions. I turned to philosophy in order to answer questions about psychiatric conditions that I couldn't answer with neuroscience alone.

4 To clarify some terminology: the brain is composed of brain states, brain properties, and brain processes, including the activity of different parts of the brain, the size and shape of different brain regions, the size and shape of nerve cells (called "neurons"), the quantities of certain proteins, the levels of certain chemicals (like neurotransmitters), the number of synapses (the spaces between individual nerve cells where communication between them occurs), etc. I will interchangeably use the terms "brain state," "brain difference," and "brain feature" to refer to these and other aspects of the brain. Furthermore, I will use "chemical difference" to refer to features of an individual's or group's chemistry (e.g., their level of neurotransmitter X in Brain region Y). Likewise, the mind is composed of mental states (examples include emotions, moods, beliefs, desires, memories, thoughts, and intentions) and mental processes (examples include calculating an addition problem; deliberating about who to ask to the prom). However, I will interchangeably use the terms "mental state," "mental difference," and/or "mental feature" to refer to any of the above. Thus "brain difference," "chemical difference,"

and "mental difference" simply refer to ways in which the brain, chemistry, or mind of an individual or group stands out from a comparison individual or group; in this chapter, these terms are not restricted to only those states that are associated with psychiatric conditions.

5 What would count as a "simplistic chemical imbalance theory"? Prime examples are those illustrated by the quotation in the introduction and those in endnote 2 (Amen 1998; Spiegel 2012; Zoloft 2001)—e.g., the theory that depression is *fully* characterized by "low serotonin in the brain."

6 Here are two more quotations:

1 From a textbook on the neuroscience of psychiatry: "Studies of neurotransmitter levels in the plasma, cerebrospinal fluid, and brain tissue have failed to find deficiencies in patients who are depressed compared with healthy controls" (Higgins and George 2013: 252).

2 According to Bruce Levin, a clinical psychologist: "Truly well-informed psychiatrists have long known that research showed that low serotonin (or other neurotransmitter) levels were not the cause of depression. The American Medical Association Essential Guide to Depression stated in 1998: 'The link between low levels of serotonin and depressive illness is unclear, as some depressed people have too much serotonin'" (Levin 2014).

See also Spiegel (2012). Furthermore, see Delgado (2000) for a review of the evidence supporting the "monoamine deficiency hypothesis" of depression (the primary "chemical imbalance theory") and the conclusion that "intensive investigation has failed to find convincing evidence of a primary dysfunction of a specific monoamine system in patients with major depressive disorders" (7). Belmaker and Agam (2008) also review evidence for the monoamine deficiency hypothesis and concludes: "Numerous studies of norepinephrine and serotonin metabolites in plasma, urine, and cerebrospinal fluid, as well as postmortem studies of the brains of patients with depression, have yet to identify the purported deficiency reliably" (56).

7 In slightly more nuanced language, there is evidence that many drug treatments (including certain antidepressants and antipsychotics) alter the balance of neurotransmitter (chemical) levels in synapses.

8 Furthermore, consider the case of antidepressants. Even though antidepressants almost certainly change a patient's neurotransmitter (chemical) balance in some way, this is likely not their ultimate therapeutic action. For example, an antidepressant like Prozac presumably increases the amount of serotonin (a neurotransmitter) in certain parts of the brain within a short time after its ingestion, but it may take anywhere from approximately two to eight weeks to actually have a therapeutic effect. So Prozac's initial increase of neurotransmitter levels, *as such*, does not seem to be the correct characterization of its ultimate therapeutic action.

9 Why do we care about "consistency," anyway? Consistency, roughly, means that we can trust that we will reliably get the same result over and over again. In other words, it means that the finding under consideration is dependable, and is not vulnerable to contradictory findings. Surprisingly, many scientists do not attempt to reproduce their own or others' findings, largely because of institutional, economic, and socio-political pressures on them to constantly publish "new" results. This is dangerous, because it means that many of the scientific results currently being reported may not be consistent, and may not hold true. A paper by Ioannidis defends this grave concern, starkly declaring that "most claimed research findings are false." (2005a: 0696; see also Ioannidis 2005b)

10 That is, findings are reproduced, but are still sometimes contradicted.

11 In general, these sorts of studies focus on two different types of brain features: (1) its function (including the relative activity levels of different parts of the brain, either when individuals are resting, or performing certain types of mental activities), and (2) its structure (including, for example, the size of certain brain regions and the density of potential connections between neurons). Both brain function and structure can be measured by different "brain scanning" technologies (often functional magnetic resonance imaging [fMRI] for function and structural magnetic resonance imaging for structure). More detailed aspects of structure (like the density of potential connections between neurons) are also sometimes assessed in postmortem samples of brain tissue. Furthermore, several reported neuroscientific results related to psychiatry are often done in animals. See Krystal and State (2014) for more information on the range of available methodologies. It is important to be aware of the limitations of these different types of methodologies. For example, regarding animal studies, it's difficult to know whether animals experience the same types of psychiatrically related mental states. And MRIs and postmortem techniques face their own unique limitations (see, for example, Dumit 2004; Lavazza and De Caro 2009; Lewis 2002; McCullumsmith and Meador-Woodruff 2011).

12 Higgins and George also mention two other regions commonly associated with depression: (1) the hypothalamic-pituitary-adrenal axis, which is a circuit existing in both the brain and body linked to stress, and (2) the nucleus accumbens, a brain region that, contrary to Higgins and George's report, is relatively absent from several analyses and meta-analyses of brain alterations in depression (see below for citations of these analyses and meta-analyses) (2013: 252–253).

13 The authors state: "Overall analyses across cognitive processing experiments ($P > .29$) and across emotional processing experiments ($P > .47$) revealed no significant results" (Müller et al. 2017: 47), and, more plainly: "neuroimaging results of UD [depression] are not consistent across experiments" (2017: 53).

14 In addition and more specifically, there are studies that contradict the prevailing wisdom mentioned above regarding *all* of the three brain regions considered to be consistently altered in depression. For example, contradictory results have been

reported in whether and how the following differ between depressed persons and neurotypicals:

- — the volume of the amygdala (Campbell et al. 2004; Price and Drevets 2009; Sheline 2003)
- — the activity of the amygdala (Gottlib and Hamilton 2008: 160)
- — the volume of the hippocampus (Campbell et al. 2004; Price and Drevets 2009; Sheline 2003)
- — the volume of the frontal cortex (Lorenzetti et al. 2009; Sheline 2003)
- — the activity of the frontal cortex (Fitzgerald et al. 2008; Gottlib and Hamilton 2008: 160; Levin et al. 2007: 218; Price and Drevet 2009: 204; Rive et al. 2013).

15 Echoing the spirit of Pauli's words, there is even inconsistency in whether authors consider findings to be inconsistent or not! For example, Gottlib and Hamilton directly contradict the above conclusions of inconsistency from Rive et al. 2013 and Fitzgerald et al. 2008 regarding the frontal cortex, declaring "remarkably consistent findings regarding abnormalities of function in this region [the DLPFC—part of the frontal cortex] have been reported" (Gottlib and Hamilton 2008: 2).

16 They are, nevertheless, important questions for other aims, such as understanding details relevant to how one *might* ameliorate or alter psychiatric conditions. See the discussions of Questions 6 and 7.

17 However, this topic has not completely vanished. Even though the direct focus of the remainder of the paper will be on neuroscientific measures of brain functioning, some of the main determinants of brain functioning are, of course, neurotransmitters and other chemicals, so the question of chemical differences/imbalances has not disappeared entirely.

18 Furthermore, we will assume a version of type physicalism that holds that all mental tokens are members of at least one mental/physical type.

19 Note, however, that token physicalism concerning the brain and mind does *not* allow for the following possibility: one token of brain state of type B is identical to a token of one type of mental state (type X) in one context, while a different token of brain state B is identical to a token of a *different* type of mental state (type Y) in another context. In other words, token physicalism *does* allow for mental states to be "*multiply realized*" by different physical tokens in different contexts, but *does not* allow for physical states to be multiply realized by different mental tokens in different contexts. See Mandik (2014) for more details.

20 Here is another way to think about it. As discussed under Section 4.1, if type physicalism is true, then *all* (types of) mental states will be identical to (types of) physical states. Love, ire, belief in God, etc. will all be identical to brain states. But, if anything is important, the experiences of love, ire, and belief in God are important, no matter what they happen to be metaphysically constructed of!

21 Or lowered, or otherwise altered. That is, just because a certain brain area is associated with anxiety, for example, doesn't necessarily mean that *more* activity in this area is identical to *more* anxiety. The relationship here could take many forms. For more about this possibility, see Hoffman and Bluhm (2016: 724).

22 One may wonder why, given the evidence of consistent brain differences discussed in Section 1.2, we are even considering the possibility of token physicalism (and the corresponding possibility that brain differences do not indicate mental ones) here. In other words, if, for example, depression is consistently associated with heightened activity in the amygdala, then why would we not assume type physicalism and say that a type of metal state (say, depressed mood) is identical to a type of brain state (say, heightened amygdala activation)? The answer follows directly from the caveats also discussed in Section 1.2. First, much of the neuroscientific evidence for a particular brain difference (or set of differences) associated with depression has been contradicted. Second, there may still be potential individual variations across any "consistent" brain differences; variations that are obscured by the fact that results are reported as averages. So even if the brain differences appear to be consistent, they may not be as consistent as type physicalism would require. Perhaps something like type physicalism roughly *appears* to be the case, but token physicalism is actually true. For example, maybe the mental state of "depressed mood" is associated with heightened activity in the amygdala in many or most individuals, but there is also a noticeable subset of individuals whose "depressed mood" is realized by a different level of amygdala activity, or by a feature of a different brain region altogether.

23 Moreover, Dr. Amen's enterprise fails to account for the possibility discussed in Section 1.2—the possibility that, even if activity of Brain Area D is *always* associated with depression in all individuals, there still may be substantial individual variations in "depressed" and "non-depressed" levels of activity in Brain Area D, such that a non-depressed person could easily possess the average "depressed activity level" in Brain Area D, and vice versa. Thus, the ultimate determinants of whether one is "really" depressed, in terms of how psychiatry currently defines "depression" in the *DSM*, will be individuals' mental states, not their brain activity. And this is arguably the case even beyond the *DSM* definition.

24 See also Hoffman (2012) and Hoffman and Bluhm (2016) for similar arguments.

25 "Underlying," here, is used as a term that is neutral between "associated with" and "identical to" in order to accommodate both dualist and physicalist positions within the philosophy of mind, respectively. The term "associated with/identical to" will also sometimes be used for the same purpose.

26 Stating it in this way papers over significant complexities. First, brain areas generally function as parts of vast networks, and may be participants in more than one network. Moreover, functioning of a brain area need not be measured exclusively in a one-dimensional range of levels of activity—there are several other

parameters, such as myriad different *patterns* of firing, along with the relative activity of specific types of neurons in each area. For example, some brain areas contain both excitatory and inhibitory neurons. Clearly, heightened activity in inhibitory neurons would presumably have a very different effect than heightened activity in excitatory neurons.

27 Although there are still serious questions about whether *seeming* to be a sign of functioning actually has any or much bearing on whether a mental state actually *is* a sign of functioning, or how "functioning" is defined in the first place.

28 Note that the same uncertainty applies to the method of determining what would count as a chemical *im*balance.

29 What is a "biological cause"? The following facts greatly impede the possibility of providing a clear definition: (1) mental states are almost certainly identical to or associated with brain (biological) states (see consideration of Question 3); (2) psychiatric conditions are incredibly complex clusters of mental and brain features that almost certainly have several causes, likely both "internal" and "external" to one's biology; and (3) there are copious interactions between the biological and the mental/psychological/social, such that any putative cause will likely produce both biological and mental/psychological/social effects, which can all serve as further causes. Psychosocial trauma, for example, can cause biological changes, and these in turn can cause or partially cause psychiatric conditions. Because of these and other factors, it seems that many causes can be described equally well as biological *and* mental/psychological/social. However, it is likely most who ask this question about "biological causes" are defining "biological causes" as initial causes that are best characterized with biological, chemical, or otherwise scientific terms. Examples include genetic variations, infectious agents like viruses, mechanical injury to the brain, exposure to toxins.

30 The only example of which I'm aware of a psychiatric condition that is regularly and exclusively caused by something biological is "general paresis of the insane" (GPI). This condition was identified in the 1800s, characterized by many mental features including delusions and mania, and later discovered to be third-stage syphilis, caused by infection with the bacterium *Treponema pallidum* (Wallis 2012).

31 For example, multiple studies show that the environment, as well as genes, plays a sizable role in the formation of (presumably all) psychiatric conditions (see Higgins and George 2013: 59–60; Plomin et al. 2016; Polderman et al. 2015).

32 Presumably along with my decision to shift my gaze. However, we will steer clear of focusing on "my decision" as a potential cause, since it opens up thorny questions about free will that are beyond the scope of this chapter.

33 See also Hoffman (2012) and Hoffman and Bluhm (2016) for similar arguments.

34 "Biological" tools, roughly, are those tools that are best referred to or named with biological, chemical, or otherwise scientific terms. Popular biological tools include

psychopharmaceuticals (drugs), electroconvulsive therapy, transcranial magnetic stimulation, and surgery. Obviously, as we have been discussing at length, this is not meant to imply that other, psychosocial tools (e.g., psychotherapy, social supports, spiritual practice) don't also harbor biological effects, but these tools themselves are arguably not best named with biological, chemical, or otherwise scientific terms.

35 Although, unsurprisingly given the discussion in Section 1.2, these studies suffer from inconsistences too: sometimes psychotherapy produces the same types of brain changes as do drugs, and sometimes it produces different types of brain changes (Barsaglini et al. 2013; Kumari 2006; Linden 2006). However, one thing that does seem to be consistent is that psychotherapy changes the brain in some way or another (Barsaglini et al. 2013; Kumari 2006; Linden 2006).

36 Nor can every type of biological input affect any type of mental change.

37 See also Hoffman (2012) and Hoffman and Bluhm (2016) for similar arguments.

38 Although it may seem indisputable, given our current societal understanding of mental differences/psychiatric conditions, that changing or eliminating such conditions would thereby better individuals' lives, this belief is not shared by several neurodivergent individuals themselves, as the Mad Pride and neurodiversity movements demonstrate.

39 Again, think of how important and how potentially liberating and/or oppressive it would be for you to regard yourself, and for others to regard you, as (1) disordered/diseased/"broken" versus (2) mentally divergent/proud. And see Hall (2016) and Nudel et al. (2009) for a small sample of examples of how various such assumptions have been liberating or oppressive for many people.

References

Amen, D. G. (1998), *Change Your Brain Change Your Life*, New York: Three Rivers Press.

Barsaglini, A., Sartori, G., Benetti, S., Pettersson-Yeo, W., and Mechellie, A. (2013), The Effects of Psychotherapy on Brain Function: A Systematic and Critical Review, *Progress in Neurobiology*, 114:1–14.

Belmaker, R. H. and Galila Agam, G. (2008), Major Depressive Disorder, *New England Journal of Medicine*, 358:55–68.

Campbell, S., Marriott, M., and Nahmias, C. (2004), Lower Hippocampal Volume in Patients Suffering from Depression: A Meta-Analysis, *American Journal of Psychiatry*, 161:598–607.

Delgado, P. L. (2000), Depression: The Case for a Monoamine Deficiency, *Journal of Clinical Psychiatry*, 61 (suppl 6):7–11.

Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., and May, A. (2004), Neuroplasticity: Changes in Grey Matter Induced by Training, *Nature*, 427:311–312.

Dumit, J. (2004), *Picturing Personhood: Brain Scans and Biomedical Identity*, Princeton, NJ: Princeton University Press.

Eklunda, A., Nicholsd, T. E., and Knutssona, H. (2016), Cluster Failure: Why fMRI Inferences for Spatial Extent Have Inflated False-Positive Rates, *PNAS*, 113:7900–7905.

Fitzgerald, P. B., Laird, A. R., Maller, J., and Daskalakis, Z. J. (2008), A Meta-Analytic Study of Changes in Brain Activation in Depression, *Human Brain Mapping*, 29:683–695.

Gotlib, I. H. and Hamilton, J. P. (2008), Neuroimaging and Depression: Current Status and Unresolved Issues, *Current Directions in Psychological Science*, 17:159–163.

Hall, W. (2016), *Outside Mental Health: Voices and Visions of Madness*, United States: Madness Radio.

Higgins, E. S. and George, M. A. (2013), *The Neuroscience of Clinical Psychiatry*, 2nd edition. Philadelphia, PA: Lippincott Williams and Wilkins.

Hoffman, G. A. (2012), What, If Anything, Can Neuroscience Tell Us about Gender Differences? pp. 30–56. In *Neurofeminism: Issues at the Intersection of Feminist Theory and Cognitive Science*. R. Bluhm, A. J. Jacobson, and H. L. Maibom, eds. Basingstoke: Palgrave Macmillan.

Hoffman, G. A. and Bluhm, R. (2016), Neurosexism and Neurofeminism, *Philosophy Compass*, 11:716–729.

Holmes, A. J. and Patrick, L. M. (2018), The Myth of Optimality in Clinical Neuroscience, *Trends in Cognitive Sciences*, 22(3):241–257.

Hoogman, M. et al. (2017), Subcortical Brain Volume Differences in Participants with Attention Deficit Hyperactivity Disorder in Children and Adults: A Cross-Sectional Mega-Analysis, *The Lancet*, 4:310–319.

Ioannidis, J. P. A. (2005a), Why Most Published Research Findings Are False, *PLoS Medicine*, 2:e124–e126.

Ioannidis, J. P. A. (2005b), Contradicted and Initially Stronger Effects in Highly Cited Clinical Research, *JAMA*, 294:218–228.

Jacobs, B., Van Praag, H., and Gage, F. (2000), Adult Brain Neurogenesis and Psychiatry: A Novel Theory of Depression, *Molecular Psychiatry*, 5:262–269.

Krystal, J. H. and State, M. W. (2014), Psychiatric Disorders: Diagnosis to Therapy, *Cell*, 157:201–214.

Kumari, V. (2006), Do Psychotherapies Produce Neurobiological Effects?, *Acta Neuropsychiatrica*, 18:61–70.

Lavazza, A. and De Carom, M. (2009), Not so Fast. On Some Bold Neuroscientific Claims Concerning Human Agency, *Neuroethics*, 3:23–41.

Levin, R. L., Heller, W., Mohanty, A., Herrington, J. D., and Miller, G. A. (2007), Cognitive Deficits in Depression and Functional Specificity of Regional Brain Activity, *Cognitive Therapy and Research*, 31:211–233.

Levine, B. (2014), Psychiatry's Manufactured Consent: Chemical Imbalance Theory and the Antidepressant Explosion, *Mad in America*. URL= https://www.madinamerica.

com/2014/05/psychiatrys-manufacture-consent-chemical-imbalance-theory-antidepressant-explosion/ Retrieved June 29, 2017.

Lewis, D. A. (2002), The Human Brain Revisited: Opportunities and Challenges in Postmortem Studies of Psychiatric Disorders, *Neuropsychopharmacology*, 26:143–154.

Linden, D. E. J. (2006), How Psychotherapy Changes the Brain—The Contribution of Functional Neuroimaging, *Molecular Psychiatry*, 11:528–538.

Lorenzetti, V., Allen, N. B., Fornito, A., and Yücel, M. (2009), Structural Brain Abnormalities in Major Depressive Disorder: A Selective Review of Recent MRI Studies, *Journal of Affective Disorders*, 117:1–17.

Maguire, E. A., Gadian, D. A., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S., and Frith, C. D. (2000), Navigation-Related Structural Change in the Hippocampi of Taxi Drivers, *Proceedings of the National Academy of Sciences*, 97:4398–4403.

Mandik, P. (2014), *This Is Philosophy of Mind*, Malden, MA: Wiley Blackwell.

McCullumsmith, R. E. and Meador-Woodruff, J. H. (2011), Novel Approaches to the Study of Postmortem Brain in Psychiatric Illness: Old Limitations and New Challenges, *Biological Psychiatry*, 69:127–133.

Ming, G. L. and Song, H. (2005), Adult Neurogenesis in the Mammalian Central Nervous System, *Annual Review of Neuroscience*, 28(1):223–250.

Moncrieff, J. (2016), Myth of the Chemical Cure, pp. 155–112. In *Outside Mental Health: Voices and Visions of Madness*. W. Hall, ed. United States: Madness Radio.

Mora, F., Segovia, G., and Del Arco, A. (2007), Aging, Plasticity and Environmental Enrichment: Structural Changes and Neurotransmitter Dynamics in Several Areas of the Brain, *Brain Research Reviews*, 55:78–88.

Müller, V. I., Cieslik, E. C., Serbanescu, I., Laird, A. R., Fox, P. T., and Eickhoff, S. B. (2017), Altered Brain Activity in Unipolar Depression Revisited, *JAMA Psychiatry*, 74:47–49.

Nudel, C., Benner, A., Poore, M., Ball, B., Knighton, D., Long, Y., Moore, K., Parrish, B., Power, C., and Ware, M., eds. (2009), *Firewalkers: Madness, Beauty, and Mystery*, Charlottesville, VA: VOCAL, Inc.

Pascual-Leone, A., Amedi, A., Fregni, F., and Merabet, L. B. (2005), The Plastic Human Brain Cortex, *Annual Review of Neuroscience*, 28(1):377–401.

Pies, R. (2011), Psychiatry's New Brain-Mind and the Legend of the "Chemical Imbalance," *Psychiatric Times*. URL= http://www.psychiatrictimes.com/blogs/couch-crisis/psychiatry-new-brain-mind-and-legend-chemical-imbalance Retrieved June 29, 2017.

Plomin, R., DeFries, J. C., Knopik, V. S., and Neiderhiser, J. M. (2016), Top 10 Replicated Findings from Behavioral Genetics, *Perspectives on Psychological Science*, 11:3–23.

Polderman, T. J. C., Benyamin, B., De Leeuw, C. A., Sullivan, P. F., Van Bochoven, A., Visscher, P. M., and Posthuma, D. (2015), Meta-Analysis of the Heritability of Human Traits Based on Fifty Years of Twin Studies, *Nature Genetics*, 47:702–709.

Price, J. L. and Drevets, W. C. (2009), Neurocircuitry of Mood Disorders, *Neuropsychopharmacology*, 35:192–216.

Rive, M. M., Van Rooijen, G., Veltman, D. J., Phillips, M. L., Schene, A. H., and Ruhé, H. G. (2013), Neural Correlates of Dysfunctional Emotion Regulation in Major Depressive Disorder. A Systematic Review of Neuroimaging Studies, *Neuroscience and Biobehavioral Reviews*, 37:2529–2553.

Sapolsky, R. M. (1996), Why Stress Is Bad for Your Brain, *Science*, 273:749–750.

Schaefer, J. D., Caspi, A., Belsky, D. W., Harrington, H., Houts, R., Horwood, L. J., Hussong, A., Ramrakha, S., Poulton, R., and Moffitt, T. E. (2017), Enduring Mental Health: Prevalence and Prediction, *Journal of Abnormal Psychology*, 126:212–224.

Sheline, Y. I. (2003), Neuroimaging Studies of Mood Disorder Effects on the Brain, *Biological Psychiatry*, 54:338–352.

Simons, P. (2018), Scientists Fight against the Myth of the Normal or Optimal Brain, *Mad in America*. URL= https://www.madinamerica.com/2018/04/scientists-fight-myth-normal-optimal-brain/ Retrieved June 14, 2018.

Spiegel, A. (2012), When It Comes to Depression, Serotonin Isn't the Whole Story, *National Public Radio Morning Edition*. URL= http://www.npr.org/sections/health-shots/2012/01/23/145525853/when-it-comes-to-depression-serotonin-isnt-the-whole-story Retrieved June 30, 2017.

Stahl, S. M. (2013), *Stahl's Essential Psychopharmacology: Neuroscientific Basis and Practical Applications*, 4th edition. Cambridge: Cambridge University Press.

Valenstein, E. S. (1998), *Blaming the Brain*, New York: The Free Press.

Wakefield, J. C. (2014), Wittgenstein's Nightmare: Why the RDoC Grid Needs a Conceptual Dimension, *World Psychiatry*, 13:38–40.

Wallis, J. (2012), Looking Back: This Fascinating and Fatal Disease, *The Psychologist*, 25:90–91.

Weinberger, D. R. and Radulescu, E. (2016), Finding the Elusive Psychiatric "Lesion" with 21st Century Neuroanatomy: A Note of Caution, *American Journal of Psychiatry*, 173:27–33.

Zoloft commercial (2001), [advertisement]. URL= https://www.youtube.com/watch?v=twhvtzd6gXA Retrieved June 29, 2017.

5

Psychopathy, Autism, and Basic Moral Emotions: Evidence for Sentimentalist Constructivism

Erick Jose Ramirez

Few human experiences are as universal as emotional experience. Or so it is thought. There are, to be sure, cross-cultural differences in emotional *concepts*. For example, the emotional life of the Ifaluk people, who all live on a few small islands within the Federated States of Micronesia, is dominated by four emotions: song, fago, ker, and metagu/rus. Although these emotional concepts are often translated as justified indignation, compassion-love-sadness, happiness-ecstasy-joy, and fear-panic-surprise-anxiety respectively, many wonder whether these translations do justice to the Ifaluk emotional life (Cardoso 2015). Although few scholars dispute the existence of cultural variation across emotional concepts, many believe that this variation masks an underlying unity of emotional *experience*.

In this chapter I provide a critical examination of what are called "basic" emotions. Philosophers and psychologists have long believed that human beings, with few exceptions, are born with the capability of feeling and expressing a small but important set of emotions. Although happiness, sadness, disgust, surprise, fear, and anger are often the prime suspects on a list of basic emotions, the exact names and number of these emotions vary widely from as few as two to as many as a eighteen (Ortony and Turner 1990).

I begin by explaining the allure that basic emotions have held for psychologists and, especially, for philosophers. Sentimentalism, the view that moral values are at least partially constituted by emotional content, lies at the heart of this allure. I will survey which, among several, senses of "basic" philosophers have traditionally been interested in when they search for basic emotions before I shift my attention to the empirical evidence for and against the existence of

these emotions. I will conclude that the best available evidence actually points us away from basic emotions and toward a conceptual act model of emotional concepts. If our interest in emotion is grounded in our interest to understand moral psychology, I argue that we can learn a lot about the role of emotion in moral psychology by examining the moral concepts of psychopaths and high-functioning persons with autism. Both psychopathy and autism are typically characterized in terms of empathic deficits. However, the empathic deficits associated with each condition are quite distinct. While individuals diagnosed with psychopathy routinely struggle to acquire moral concepts, high-functioning persons with autism do not. I conclude by suggesting the psychopath's lack of moral concepts is best understood as a result of their lack of empathic contagion and is exactly what a conceptual act model of emotion would predict.[1] Sentimentalist moral psychology, I claim, is enriched by incorporating these findings into their normative frameworks.

The allure of basic emotions

Aside from those with purely theoretical interests, why might we care about whether our emotions are basic (in some sense) or not? One answer comes to us courtesy of early modern philosopher David Hume. Hume argued that our emotions, not our reasons, are responsible for our moral and aesthetic values (Hume [1740]2002). When we judge that something is beautiful or ugly, right or wrong, just or unjust, what we are doing, Hume claimed, is projecting our emotional states onto the external world. To better understand the evaluative world, Hume argued, we must better understand our emotional lives. Hume is not alone in sensing this connection between how we *feel* about something and our value *judgments* about it. Sentimentalist philosophers believe that this observation represents a deep truth about the nature of morality.

Basic emotions are especially interesting to philosophers trying to make sense of the link between emotion and value. Sentimentalist philosophers turn to basic emotions to explain two seemingly essential features of morality. First, it seems true for most of us that moral judgments are not passive experiences. To judge that something is wrong is to feel some kind of motivation to *do something* about our judgments. If moral judgments were purely about what we *think*, this feature of our moral judgments would be harder to explain. Picture someone who believes that humans are causally responsible for global climate change. Assume further that this person does not think that it's *wrong* for humans to

cause global climate change. Would we be surprised if such a person fails to do anything about climate change? Probably not. Now picture someone who believes that anthropogenic climate change represents an injustice both to future generations and to the globe's poorest people. It would seem odd if someone who sincerely judged it wrong to contribute to climate change did nothing or, worse, increased her carbon footprint. Sincere moral judgments, in other words, seem to be at least weakly motivating. We can make sense of this fact, assuming it is a fact, if our sincere moral judgments were emotionally laden. Emotions are intrinsically motivating in the sense that feeling an emotion brings with it an impulse to act.

Second, emotions appear to lie at the bottom of many moral and aesthetic disagreements. In a series of studies, psychologist Jonathan Haidt asked subjects whether certain actions were wrong. These included instances of harmless but clearly aberrant behavior. For example, Haidt included cases where subjects used an American flag to clean a toilet or in which a family decides to eat their recently deceased dog instead of burying it (Haidt, Helena Koller, and Dias 1993). In another study, Haidt asked his subject to determine whether or not it would be wrong for two siblings to engage in a consensual incestual relationship (Haidt 2001). As a result of a phenomenon now referred to as "moral dumbfounding," Haidt concluded that his subjects grounded their judgments about the moral permissibility of an act based not on good reasons but instead on how those acts made them *feel*. For example, Haidt found that his subjects continued to say that consensual incest was wrong even when he concocted cases in which all of the reasons they gave to justify their judgments no longer applied.[2] "Moral dumbfounding" remains a controversial thesis about evaluative judgment (Jacobson 2013). However, sentimentalist philosophers see it as powerful evidence in favor of their position. To these philosophers, emotions appear to play a fundamental role in our moral judgments and moral practices.

If our moral judgments are, in fact, grounded on emotion, then the allure of basic emotions is easier to understand. Basic emotions offer a solution to a problem that many philosophers believe must be explained by any adequate moral theory: the problem of disagreement. It is a truism to claim that people disagree about values. It is harder to explain what these disagreements say about the nature of morality.

Disagreements, for example, occur both within a culture and between cultures. Whether abortion ought to be permissible, for example, continues to be the subject of intense disagreement within and between cultures. Although many disagreements might be explained by the fact that one, or both, parties to

the disagreement have imperfect or incorrect information, it seems possible that equally well-informed persons can disagree about what goodness consists in or whether a particular action is morally justifiable. Although many philosophers confront moral disagreement by embracing moral relativism (the view that moral claims are limited in their scope or application), sentimentalists hoping for a more objective response to the problem of moral disagreement have turned to basic emotions for help.

> The idea is that a shared response, or sentiment, somehow moors us in a common subject matter, making it possible for us to disagree substantively about what a thing has to be like in order to be such that we should feel this sentiment toward it. Thus, if the sentimentalist is right, it is because our evaluative concepts have a special tie to shared human sentiments that we are able to engage meaningfully in debates over their application. (D'Arms 2005: 13)

If evaluative judgments are anchored by our emotional responses and at least some of those emotional responses are basic, then moral disagreements can be understood as disagreements about the conditions under which a particular emotion should be felt. If these emotions are *basic* then there also may be a mind-independent answer available to help settle such disputes. In other words, basic emotions offer the prospect of securing a universal domain of cross-cultural human values. To see why such objective standards exist, I need to say more about what senses of "basic" philosophers and psychologists have in mind when they speak of "basic emotions."

What's basic about basic emotions?

What are we saying when we say that an emotion is basic? In this section I briefly survey three possible answers to this question. Emotions might be basic in a developmental, compositional, or phylogenetic sense (Ortony and Turner 1990). Importantly, an emotion may be basic in all three senses. They are not mutually exclusive. Although it is a live question whether emotions are basic in any sense, in this section I argue that phylogenetically basic emotions hold the greatest promise for sentimentalist moral philosophers.

To say that an emotion is developmentally basic is to say that the emotion appears early in the course of human development. That some emotions may be developmentally basic is not particularly debated among philosophers or psychologists. Prime candidates for developmentally basic "emotions" include positive and negative feeling states. These states define the valence of all

experience. Although positive and negative feelings are an important part of lived experience, such states are too "thin" to serve the philosophically complex roles that moral emotions do.[3] They cannot ground moral disagreement and lack the content to explain the differences between different kinds of moral assessments.

To say that an emotion is compositionally basic is to say that these emotions work like atoms with respect to molecules. Compositionally basic emotions can be used to build other emotions and do not contain any other emotions as parts.

> The main criterion used by those who advocate basic emotions in the [compositional] sense focuses on the interrelationship of the emotions, rather than directly on the nature of the eliciting conditions. Here, an emotion is regarded as basic if it contains no other emotion as a component. (Ortony and Turner 1990: 324)

Schadenfreude, for example, is the name given to a type of pleasure derived from the suffering of others. Schadenfreude is unlikely to be compositionally basic. It decomposes into several more primitive states: empathic pain derived from the suffering of others and pleasure caused by that experience. Many emotions might be thought to be compositionally basic though this form of basicness also will not be sufficient to serve the philosophically useful functions that philosophers appeal to basic emotions for. Compositionally basic emotions may vary widely across cultures and hence fail to account for disagreements.

Lastly we can call an emotion basic in a phylogenetic sense. To call an emotion phylogenetically basic is to say something about the *causal* history of that emotion. A phylogenetically basic emotion arises as a product of natural selection. Psychologist Paul Ekman writes that:

> [t]here are two key issues, which I use the adjective *basic* to convey about the position I have adopted … (1) There are a number of separate emotions which differ one from another in important ways. (2) Evolution played an important role in shaping both the unique and the common features which these emotions display as well as their current function. (Ekman 1994: 170)

Prime candidates for phylogenetically basic emotions include anger, disgust, fear, surprise, happiness, and sadness. A phylogenetically basic emotion is also likely compositionally basic in the sense that human cultures can expand upon their evolutionarily gifted repertoire of emotions to form more complex blends (transforming guilt into survivor's-guilt, for example). That an emotion is phylogenetically basic therefore makes it a prime candidate for solving the problems of moral disagreement and for making sense of Haidt's cross-cultural moral dumbfounding research.

This is because if phylogenetically basic emotions exist then all neurotypical humans will share these emotions. If all neurotypical humans share these emotions then they can serve as common points of contact for moral conversations about the function and purpose of these emotions. Additionally, because they are evolutionary products, there will be an answer to questions about what the appropriate objects are for these emotions. This would allow empirically inclined philosophers to make progress on addressing difficult moral questions that are grounded on those emotions. For example, if shame evolved to serve a specific purpose then there may be a way to answer questions about whether it is appropriate to be ashamed of our appearance, class standing, or gender identities. It is little wonder that sentimentalist philosophers have tended to focus their attention on phylogenetically basic emotions (D'Arms and Jacobson 2000). In what follows, I use the term "basic emotion" to refer to phylogenetically basic emotions.

Evidence for basic emotions

What evidence do we have that emotions are basic? Charles Darwin theorized that animals (including human animals) behaved in relatively stereotyped ways when placed in specific sorts of circumstances (Darwin [1873]2009). Dogs will often adopt threatening or submissive body postures, for example, to establish a pack hierarchy. Humans will often flush and hide their faces when performing in socially costly ways. Emotions, he reasoned may have evolved as strategic solutions to universal life problems and his observations of stereotyped, seemingly automatic, behavior was evidence of this. In the wake of Darwin's arguments, behavioral evidence dominated in the search for basic emotions.

In this section, I survey this evidence. I focus predominantly on facial expression research that has been emerging out of Paul Ekman's research group for over four decades. I argue that, while Ekman's research suggests that emotion and facial expression are connected, facial expressions do not give us evidence for basic emotion. Some have also looked to the body and the brain as a source of evidence for basic emotions. Although many have claimed to find evidence for emotionally distinct physiological or neurological profiles, I argue that meta-analyses of these studies provide, at best, evidence for a much weaker thesis than the existence of basic emotion. Evidence about a person's face, body, and brain can give us useful information about the valence of their experience and about its intensity but no combination of these three pieces of information will tell

us which particular emotion someone is experiencing. We will need to turn to other models of emotion to explain this result.

In 1971, Paul Ekman and Wallace Friesen published a study on facial expressions that would create a new paradigm in emotion research (Ekman and Friesen 1971). Ekman and Friesen believed that facial expressions represented specific emotional states. To test this claim, Ekman and Friesen needed to determine whether people the world over agreed about what facial expressions meant.

Ekman and Friesen chose the Fore people of Papua New Guinea as subjects. The Fore were chosen because they lived in small groups that were relatively isolated from Western cultural influences. Ekman and Friesen showed pictures of American subjects to members of the Fore. Given their isolation, the Fore were seen as excellent subjects to test Darwin's supposition about the link between behavior, especially facial expression, and emotion. Each picture was a closely cropped image of a person's face which Ekman and Friesen had taken great pains to pose. Each Fore was presented with one of three emotion words (translated from English into the Fore's native language) and then asked which emotion the face in the image represented. If facial expressions gained their meaning a result of socio-historical facts then the Fore, who lacked knowledge of Western culture, should have responded somewhat randomly when asked to match facial expressions to emotion words. This is not what Ekman and Friesen observed. The Fore converged on the same sorts of answers that their Western counterparts converged on. Smiling faces were connected to happiness, frowning faces with anger, and so on.

Other researchers would quickly adopt Ekman and Friesen's methodological approach and apply it across the globe (Ekman 1972b; Matsumoto 1992). The seemingly stunning convergence of the world's peoples to questions about the meaning of facial expressions was interpreted as a powerful vindication of Darwin's suggestion that emotions were basic. Ekman argued that basic emotions were what he called "affect programs." Affect programs are mental modules. As such, they are non-conscious evolved neural circuits which produce stereotyped responses. Affect programs are triggered by specific emotionally relevant input and produce emotionally distinct outputs (Ekman 1972a). Each emotion, on this view, is distinguishable from others in virtue of the inputs and outputs associated with it. Fear, for example, can be understood as a response to the perception of something as an imminent threat. When triggered, the fear module activates and prepares a creature to fight or flee the source of the threat. These preparations include emotion-specific changes to a creature's physiology

(increased heart rate, blood pressure, and so on) and, in humans and other apes, includes the production of an emotion-specific facial expression. Although the facial expression research paradigm has been successfully replicated by researchers for decades, we should be wary of the conclusion that Ekman and others have drawn from it. In particular, I suggest that we jump too hastily from the convergence of facial expression judgment studies to claims about affect programs and basic emotion (Russell 1995).

The affect program approach to emotion is grounded on several assumptions. First, the claim that emotions evolved to produce specific facial expressions. Second, that humans evolved to detect these signals to better understand the emotional lives of those around us. Third, that these signals are difficult to convincingly fake. All three assumptions have been challenged in the last twenty-five years.

Facial expression judgment studies have been subject to many lines of criticism. One line of criticism takes issue with Ekman's methodology, which most facial expression judgment studies adopted. Ekman's paradigm utilized what is known as a within-subjects forced-choice design. Although, by itself, the use of any particular study design is neither helpful nor hurtful, methodological decisions can have an impact how on we analyze the results of each study (Russell 2016). A "forced-choice" experiment is one in which subjects must select a response from those provided by the research. In the case of facial expression studies, critics point out that the forced-choice response format artificially raises the degree of convergence. The problem with using forced choice within subject experiments was perhaps best demonstrated in a series of studies in which, depending on the options available, subjects could be made to converge on different answers, even nonsense answers, when asked what a specific face expresses (DiGirolamo and Russell 2017).

To say that an experiment is "within-subject" is to say that each subject of the experiment is exposed to the entire stimulus set (Russell1995). In the context of facial expression studies, it means each subject is shown all of the facial expression photographs used in the experiment and asked to make judgments about them. Why might this pose a problem for judgment studies? Repeated exposure to several faces cues subjects to pay closer attention to the differences between faces and therefore impacts their judgments about the meaning of each face. This feature serves to artificially raise rates of convergence. Facial expression judgment studies that adopt "free-response" formats (where subjects are asked to write in what they believe a face expresses) or "between-subject" designs (where subjects are only shown a single facial expression) show significantly lower rates

of convergence than Ekman's more traditional protocol. This is especially true for studies using members of non-Western cultures as subjects (Gendron et al. 2014).

Critics argue that these critical experiments show us that it is the study *design* utilized by most facial expression researchers that produces such high rates of convergence on answers. While these methodological criticisms do not rule out the existence of basic emotions, they weaken the kind of evidence that facial expression studies provide for them.

A second set of criticisms levied against the facial expression paradigm targeted the assumption that basic emotions evolved to produce specific facial expressions as signals. In a study, conducted in the 1920s, psychologist Carney Landis tried to trigger strong emotional states in his subjects. Landis did this in various ways including asking subjects to place their hands into a bucket of frogs, smell ammonia, and view at pornographic images; at the end of the experiment, Landis asked his subjects to decapitate a living rat (Landis 1924). Throughout the experiment, Landis took photographs of his subject's facial expressions. To his surprise, he found that subjects rarely produced the stereotyped facial expressions commonly associated with basic emotions. Although Landis's approach has methodological (and ethical) issues of its own, Landis's data would prompt other researchers to test the assumption that basic emotions evolved to produce specific facial expressions.

Jose Fernández-Dols and his colleagues would attempt to test these assumptions in the late 1990s. They were interested in the kinds of facial expressions people spontaneously made while undergoing strong emotional episodes. For example, they observed gold medal winners when they won their competitions, as they waited during the award ceremony, and during the moment they had their medals placed around their necks (Fernández-Dols and Ruiz-Belda 1997). They also studied facial expressions generated spontaneously in a number of different public spaces including those of soccer fans during a match (Fernández-Dols and Crivelli 2013). If facial expressions evolved as non-cognitive (unconscious) signals of basic emotions, then soccer fans and gold medal winners seem like the perfect test subjects.

Like Landis, Fernández-Dols's results did not conform to the expectations of the facial expression paradigm. He discovered that, while subjects reported strong feelings throughout their gold medal experience, for example, that they only produced the expected facial expression when they were under public scrutiny. As the medal was placed on their necks, gold medal winners smiled. Smiles were missing, however, during other elements of their experience.

Fernández-Dols and others argued that facial expressions were more likely to be the product of social signaling than of automatic affect programs. Attempts to discover the strength of the link between felt emotion and facial expression has consistently demonstrated that link to be weaker than predicted by affect program theorists (Reisenzein et al. 2013). Coupled with the methodological criticisms of the affect program paradigm, emotion researchers began to look for other avenues in support of basic emotions.

Evidence from facial expression judgment studies, while initially seen as strongly supporting basic emotions, has received a critical second-look in the late twentieth and early twenty-first centuries. Basic emotion theorists have, in the meantime, tried to look for other types of evidence to support their views. In particular, these researchers have held out hope that our bodies and brains will provide them with tell tale clues of basic emotions. If discrete emotions evolved to solve specific life conflicts, these emotions might be identifiable by the discrete patterns of physiological or neurological activation they trigger. An explosion of studies emerged in the late 1990s and 2000s all claiming to have discovered emotion-specific markers for anxiety, anger, sadness, disgust, happiness, and the like. However, when subjected to meta-analysis, the strength and uniqueness of these emotion-specific markers falls dramatically (Phan et al. 2002; Wager et al. 2015). At the time of writing, intense disagreement exists about the robustness of emotion-specific physiological or neurological markers.

Sentimentalist philosophers appealed to basic emotions in the hopes of addressing long-standing questions about the nature of moral judgments and moral disagreements. Basic emotions seemed poised to explain how moral disagreements, even radical ones, could be understood without appeal to moral relativism. However, in light of what has become a growing consensus away from basic emotions among psychologists, neuroscientists, and philosophers of mind, sentimentalists must look elsewhere for means of addressing these concerns. The picture of emotions emerging leans away from emotions as basic and toward a view of emotions as culturally informed perceptions.

Beyond basic emotions

Although we have reason to doubt the existence of basic emotions, sentimentalist philosophers can salvage a great deal of useful material from the theories of emotion emerging in response to the critiques discussed above. In addition, recent discoveries about the varying capacities all currently referred to as forms

of empathy, especially as they appear in psychopaths and high-functioning people with autism, not only give further evidence for the sentimentalist approach but can functionally replace basic emotions. That is, constructivist accounts of emotion can fulfill most of the roles that basic emotions were once thought uniquely suited perform. In this section I briefly survey contemporary psychological constructivist and conceptual act theories of emotion. In the following, and final, section, I explain how psychopathic and autistic agency can point the way toward a more satisfactory moral psychology.

If emotions are not basic, then what are they? How do we explain the seeming universality of emotional experiences and emotional facial expressions? Psychological constructivist approaches to emotion address these questions by distinguishing between "affect" and "emotion." Affect, on this view is a phylogenetically basic aspect of all experience including non-emotional experiences. Affect defines the "feel" or "phenomenology" of experience and is defined in terms of valence (positive/negative) and arousal (high/low) (Feldman Barrett 2006). Hunger, for example, might include a highly arousing but negative affective component. Although emotions include affect as a component, emotions are meaningful groupings of causes, affective changes, thoughts, and behaviors that have been given their meaning as a result of cultural construction.

> Like the basic emotion approach, psychological constructionist models take evolution seriously and consider how biological and species-general factors have some role in the events that are conventionally named as "'sadness'", "'anger'", and "'fear'". Like the appraisal approach, the psychological construction approach to emotion involves making meaning out of sensory cues. And, like social constructionist models of emotion, psychological construction approaches consider the [culturally] specific elements of emotion. As socially constituted artifacts of learning and culture, no emotion category is assumed to be biologically basic. (Barrett 2009: 1290)

Emotional concepts have top-down effects on affective experience. They structure how we understand the causes of affective change and the meaning of the change itself. Fago, song, and the like are a part of the emotional and social world for the Ifaluk. Emotional concepts structure and regulate how a person sees the world. In much the same way that we do not have to consciously make efforts to parse out our perceptions *into* cars, airplanes, computers, and the like, our emotional concepts structure our affective experiences relatively automatically. Because emotional concepts must be learned (they are language and culture dependent), these approaches to emotion are comfortable abandoning the basic emotion paradigm.

> According to the conceptual act model, emotion categories exist because groups of people agreed (for phenomenological and social reasons) that this is a functional way to parse the on-going mental activity that is realised in the brain. The model is consistent with the observation that some of the categories are cross-culturally stable (because they function to address certain universal human concerns that stem from living in large, complex groups), whereas other categories are culture specific. The conceptual act model hypothesises that the category instances named with emotion words are real, but they derive their reality from the human mind (in the context of other human minds). (Barrett 2009: 1293)

Conceptual act theories allow for a great deal of variation in terms of how cultures parse out affective space into emotional and non-emotional experiences. For example, "[m]any depressed Chinese people do not report feeling sad, but rather express boredom, discomfort, feelings of inner pressure, and symptoms of pain, dizziness, and fatigue. These culturally coded symptoms may confound diagnosis among Chinese immigrants in the United States, many of whom find the diagnosis of depression morally unacceptable and experientially meaningless" (Kleinman 2004). Such findings are consistent with the conceptual act theory.

I pause here to note that, on the conceptual act model of emotion, emotions are cultural artifacts. Sentimentalists interested in grounding moral values in emotion will therefore be driven to adopt a form of cultural relativism. Moralized emotional concepts will only make sense within those groups where such an emotion is available to structure affective experience. However, I argue that this acceptance has two upsides.

First, at least some emotional concepts are likely to be pancultural in virtue of being useful for structuring social relations and managing the distribution of scarce resources. Shame may be such an emotion (Wong and Tsai 2007). Insofar as moral regulative emotions like shame are cross-culturally shared, their lack of basicness does not pose a problem for cross-cultural moral communication. Second, conceptual act models make predictions about how individuals acquire and use emotionally laden moral concepts. In the next section, I argue that investigations of the moral concepts of psychopathic and high-functioning persons with autism largely support these predictions and provide further insight into the nature of our moral capacities.

Psychopathy, autism, and sentimentalist moral psychology

Conceptual act theories of emotion include a basic element, affect, and a constructed element, emotion. Although affect is a complex phenomenon, some

elements may be more important toward the acquisition of moral concepts than others. In particular, although psychopathy and autism are often characterized in terms of empathic deficits, I argue that these deficits are radically different from one another. High-functioning people with autism appear capable of acquiring moral concepts while psychopaths cannot. An investigation into their empathic capacities will help us make sense of our underlying moral capacities.

Psychopathy is a controversial name for a controversial diagnosis. In referring to psychopathy, I refer to the condition first identified by Hervey Cleckey in the 1940s and whose diagnostic criteria have been refined by Robert Hare and his psychopathy checklist (PCL-R). I do not simultaneously intend to refer to the diagnosis of antisocial personality disorder (ASPD) in the 5th edition of the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association 2013; Hare et al. 1991). Psychopathy, on this understanding, names a complex spectrum disorder diagnosed primarily in terms of character traits and emotional capacities. Because it is a spectrum disorder, I limit my attention to a subclass of psychopaths referred to as successful psychopaths. Successful psychopaths score highly on the PCL-R but also lack other rationality-undermining comorbid illnesses (Babiak et al. 2010).

Psychopaths are said to lack an ability to feel empathic distress. Empathic distress occurs when we find ourselves immediately pained when we witness another person suffering. It is the name given to a particular form of empathy that is basic in all three senses of the term. It is developmentally basic in the sense that, for nearly all humans, empathic distress develops by around twenty-four months. It is compositionally basic in the sense that empathic distress cannot be broken down into still more primitive experience. Lastly, because this capacity is found in many mammalian species, there is evidence that it is phylogenetically basic as well. Empathic distress is thought to be mediated by mirror neurons in somatosensory cortex that activate in response to our perception of the suffering of others (Fecteau et al. 2008). In addition to lacking empathic distress, psychopaths are also characterized by personality traits including glibness, superficial charm, deceitfulness, and so on (Ramirez 2016). Although psychopaths are characterized in terms of an inability to feel empathic distress, successful psychopaths are often masters of another form of empathy: mindreading. This form of "empathy is the process by which one attains a cognitive grasp of, belief about, or knowledge of another's mental states" (Battaly 2011: 286). Mindreading empathy is an especially useful trait for the psychopath as it allows them to deceive and manipulate others by predicting their mental states.

Autism is also a spectrum disorder. Autism, like psychopathy, is often characterized in terms of emotional deficits, empathic difficulty, and impairments

in social function (McIntosh et al. 2006).[4] Although both psychopathy and autism include empathic deficits, these deficits refer to different capacities. Where the psychopath lacks a capacity for empathic distress and excels at mindreading, the high-functioning agent with autism has an intact capacity for empathic distress but struggles with basic mindreading abilities.[5]

The empathic differences between high-functioning persons with autism and psychopaths translate into differences in moral understanding and moral capacity. Psychopaths have difficulty distinguishing between what psychologists call moral norm violations and conventional norm violations. Although the distinction is controversial, norms of convention are thought to be authority-dependent, not based on harm, and limited in their context. For example, the norm for raising one's hand before speaking is authority-dependent. If a speaker announces that her audience should feel free to ask questions without raising their hands, the rule no longer applies. When studying these distinctions, psychologists construe moral norms as authority-independent, serious, harm based, and universal.[6] Psychopaths tend to see all norms as norms of convention (Dolan and Fullam 2010). Additionally, psychopaths lack the sort of moral motivation I referred to earlier when discussing the special nature of sincere moral judgments. Because they see all norms as mere norms of convention, moral norms lack special force.

Most of us can easily note, for example, that there is a difference between how norms like "it is wrong to wear white after Labor Day" and "it is wrong to steal a person's identity." Whereas the powers that be may change fashion norms, prohibitions against harming others non-consensually would continue to apply even if an authority figure claimed that they did not. The tendency to see moral norms as independent from authority runs so deep that many, including religiously orthodox communities, see moral norms as independent of god's authority (Nucci and Turiel 1993).

Children, including children with autism, begin to mark the moral/conventional distinction at a little over two years of age (Turiel 1977). Psychopaths, on the other hand, do not consistently or clearly note the difference. Although the moral concepts of individuals with autism can be structured differently from neurotypical moral concepts, their ability to mark the moral/conventional distinction and, especially, their desire to universalize the application of moral rules, makes clear that the moral concepts of persons with autism are full-blown moral concepts (Kennett 2002; McGeer 2008). The conceptual act theory of emotion can help to explain why persons with autism retain their capacity to form moral concepts and why psychopaths cannot.

The conceptual act theory of emotion states that emotional concepts are acquired via cultural transmission. The primary vehicle for the cultural transmission of moral concepts is empathic contagion (Ramirez 2016). In particular, children with an intact capacity for empathic contagion will mimic the emotional expressions of those around them, learn to give specific patterns of emotional expressions names, and thereby develop emotional concepts that structure their adult affective experiences (Feldman and Barrett 2006). Different cultures will structure their emotional concepts differently, according to their own needs, though some particular emotional concepts will prove cross-culturally useful.

If moral concepts gain their meaning from emotional concepts, as sentimentalists claim, then empathic contagion plays an essential role in the acquisition of moral concepts. Because high-functioning persons with autism are capable of empathic distress, they will still be able to form moral concepts. Their issues with mindreading can make the application of these concepts difficult but their ability to note the distinction between purely authority-driven norms of convention and harm-based moral norms make it clear that they understand the moral domain as a distinct, affectively backed, domain (McGeer 2008).

Psychopaths, on the other hand, lack the capacity to engage in empathic distress. They thus cannot acquire moral concepts using the traditional route. Their moralized concepts have the form of descriptive anthropological data. Their inability to feel pained by the harm they cause or to feel guilt as a result of violating moral wrongs can thus be traced back to their inability to form genuine moral concepts. This is just as the conceptual act theory of emotion predicts.[7]

Sentimentalists interested in acquiring an empirically adequate picture of the nature of moral concepts and the structure of moral education ought to pay attention to the moral concepts of high-functioning persons with autism and of psychopaths. From these groups we learn that empathic contagion is the primary vehicle for the acquisition of moral concepts. Given the constructivist nature of the conceptual act theory, a sentimentalist constructivism emerges. Although this might appear to cut off the possibility of moral univocity, it does no such thing. Agricultural practices are fairly common, for example, and fairly stereotyped. The importance of crop rotation, irrigation, and other practices has been independently discovered several times in different cultures. Agriculture is not a basic human practice, however. Similarly, emotions are not basic in the senses mentioned earlier in this chapter (though our tendency to form emotional concepts may be). Some emotions may be so useful, however, that many cultures may construct emotions that all bear a family-resemblance to one

another. Shame, for example, is an intensely social emotion and can track social relationships in highly complex ways. In cultures where such relationships are highly valued, an emotion very much like shame is likely to be found. Because at least some emotional concepts are useful to regulate human social interactions (shame and embarrassment being prime candidates), these emotions can still serve as a basis for explaining moral disagreements in a more limited fashion. To the degree that emotional concepts across cultures overlap with one another, then evaluative discourse about the values these emotions carve out is possible. Though less far-reaching than basic emotions, these emotions would permit cross-cultural moral engagement.

Conclusion

Sentimentalist philosophers have turned to psychology and the neurosciences in the hopes of vindicating their intuition that moral concepts are grounded in our emotions. Though initially this empirical turn appeared to support the existence of basic emotion, these data have received significant criticism. In place of basic emotion, conceptual act theories have emerged to explain the seeming variation and universality of emotional experience. These views receive further support from the moral concepts of high-functioning persons with autism and psychopaths. These groups provide strong evidence for the claim that empathic contagion is the primary vehicle for the acquisition of moral concepts. Because empathic contagion is basic, we should expect to see this capacity in all neurotypical human populations. Sentimentalists worried about the prospects of constructivism that emerges from conceptual act approaches can manage to explain many moral disagreements even without appeal to basic emotion.

Notes

1 Throughout this chapter, I make reference to "psychopaths." Human beings are neurodiverse, meaning that human beings experience and cognize the world and human social concepts in diverse ways. In using the term psychopathy, I intend to describe a class of persons diagnosed with a very specific set of character traits (elaborated in Section 5 of this chapter). I do not intend to license any normative judgments about the status, standing, or value of such persons.

2 The most common reasons given against incest included the risk of birth defects and a concern over possible abuse. In one version of Haidt's incest case,

for example, the siblings met only as adults, did not have a history of abusive relationships, highly enjoyed the experience, and agreed to be sterilized before sexual contact. In such cases, subjects continued to judge that the act was wrong. When pressed for a reason, many claimed that incest was simply disgusting.

3 To call an emotion "thin" is to say that it lacks evaluative content. To say that a state feels good (or bad) is not to say all that much about it. Contrast that with saying that one is disgusted by something. Disgust is "thicker" in that it not only tells us something about the *valence* of the feeling, it *condemns* its object.

4 Proponents of neurodiversity would, rightly, reject characterizations of autism that cast it as a mental illness or disorder (Ramirez 2016). For the purposes of my argument, I focus only on population-level differences between persons with autism, psychopaths, and neurotypical persons. I leave aside questions about whether autism is best viewed as a mental disorder or as an example of variation in human modes of cognizing the world.

5 As with any spectrum disorder, generalities like these paint very broad strokes about the capacities that individuals have. Although some researchers have argued that persons with autism do in fact have a diminished capacity for empathic distress, in comparison with the deficits seen with psychopaths, the differences between high-functioning persons with autism and neurotypical persons are less stark. For example, psychopaths have profoundly diminished myelination in areas thought to contain mirror neurons which might explain their pronounced lack of empathic distress (Ly et al. 2012).

6 Although the moral/conventional distinction is controversial, I use it here to show that psychopaths inherit a kind of moral disability as a result of their lack of empathic distress (Greenspan 2003). I need not commit myself to the claim that the sense of "moral" defined by the moral/conventional task is essentially correct.

7 Psychopaths are not entirely devoid of moral emotions. Psychopaths are capable of intense anger, for example and so-called "secondary" psychopaths appear capable of a degree of shame though this shame is often focused on failing to achieve the recognition that secondary psychopaths believe that they are owed (Morrison and Gilbert 2001).

References

American Psychiatric Association (2013), *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Arlington, VA: American Psychiatric Publishing.

Babiak, P., Neumann, C., and Hare, R. (2010), Corporate Psychopathy: Talking the Walk, *Behavioral Sciences and the Law*, 28:174–193.

Battaly, H. D. (2011), Is Empathy a Virtue? In *Empathy: Philosophical and Psychological Perspectives*. A. Coplan and P. Goldie, eds. Oxford: Oxford University Press.

Cardoso, F. M. S. (2015), Social Regulation of Emotion: A Foray into the Work of Catherine Lutz, *Revista E-Psi*, 5:103–109.

D'Arms, J. (2005), Two Arguments for Sentimentalism, *Philosophical Issues*, 15:1–21.

D'Arms, J. and Jacobson, D. (2000), Sentiment and Value, *Ethics*, 110:722–748.

Darwin, C. ([1873] 2009), *The Expression of the Emotions in Man and Animals*, Oxford: Oxford University Press.

DiGirolamo, M. A. and Russell, J. A. (2017), The Emotion Seen in a Face Can Be a Methodological Artifact: The Process of Elimination Hypothesis, *Emotion*, 17: 538–546.

Dolan, M. and Fullam, R. (2010), Moral/Conventional Transgression Distinction and Psychopathy in Conduct Disordered Adolescent Offenders, *Personality and Individual Differences*, 49:995–1000.

Ekman, P. (1972a), *Emotions in the Human Face*, New York: Pergamon Press.

Ekman, P. (1972b), Universal and Cultural Differences in Facial Expressions of Emotions, pp. 20–283. In *Nebraska Symposium on Motivation*. J. K. Cole, ed. Lincoln: University of Nebraska Press.

Ekman, P. (1994), An Argument for Basic Emotions, *Cognition and Emotion*, 6:169–200.

Ekman, P. and Friesen, W. (1971), Constants across Culture in the Face and Emotion, *Journal of Personality and Social Psychology*, 17:124–129.

Fecteau, S., Pascual-Leone, A., and Theoret, H. (2008), Psychopathy and the Mirror Neuron System: Preliminary Findings from a Non-Psychiatric Sample, *Psychiatry Research*, 160:137–144.

Feldman Barrett, L. (2006), Solving the Emotion Paradox: Categorization and the Experience of Emotion, *Personality and Social Psychology Review*, 10:20–46.

Feldman Barrett, L. (2009), Variety Is the Spice of Life: A Psychological Construction Approach to Understanding Variability in Emotion, *Cognition & Emotion*, 23:1284–1306.

Fernández-Dols, J. M. and Crivelli C. (2013), Emotion and Expression: Naturalistic Studies, *Emotion Review*, 5:24–29.

Fernández-Dols, J. M. and Ruiz-Belda, M. A. (1997), Spontaneous Facial Behavior During Intense Emotional Episodes: Artistic Truth and Optical Truth, pp. 255–274. In *The Psychology of Facial Expression*. J. A. Russell and J. M. Fernández-Dols, eds. New York: Cambridge University Press.

Gendron M., Roberson D., Van Der Vyver, J. M., and Feldman Barrett, L. (2014), Perceptions of Emotion from Facial Expressions Are Not Culturally Universal: Evidence from a Remote Culture, *Emotion*, 14:251–262.

Greenspan, P. S. (2003), Responsible Psychopaths, *Philosophical Psychology*, 16:417–429.

Haidt, J. (2001), The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment, *Psychological Review*, 108:814–834.

Haidt, J., Helena Koller, S., and Dias, M. G. (1993), Affect Culture and Morality, or, Is It Wrong to Eat Your Dog?, *Journal of Personality and Social Psychology*, 65:613–628.

Hare, R., Hart, S., and Harpur, T. (1991), Psychopathy and the DSM-IV Criteria for Antisocial Personality Disorder, *Journal of Abnormal Psychology*, 100:391–398.

Hume, D. ([1740] 2002), II.3.3 415. In *A Treatise of Human Nature*. D. F. Norton and M. J. Norton, eds. New York: Oxford University Press.

Jacobson, D. (2013), Moral Dumbfounding and Moral Stupefaction, pp. 289–316. In *Oxford Studies in Normative Ethics*. M. Timmons, ed. Oxford: Oxford University Press.

Kennett, J. (2002), Autism, Empathy and Moral Agency, *The Philosophical Quarterly*, 52:340–357.

Kleinman, A. (2004), Culture and Depression, *New England Journal of Medicine*, 351:951–953.

Landis, C. (1924), Studies of Emotional Reactions: General Behavior and Facial Expression, *Journal of Comparative Psychology*, 4:447–509.

Ly, M., Motzkin, J. C., Philippi, C. L., Kirk, G. R. Newman, J. P., Kiehl, K. A., and Koenigs, M. (2012), Cortical Thinning in Psychopathy, *The American Journal of Psychiatry*, 169:743–749.

Matsumoto, D. (1992), More Evidence for the Universality of a Contempt Expression, *Motivation and Emotion*, 16:363–368.

McGeer, V. (2008), The Makings of a Moral Sensibility: Replies to Commentaries, pp. 281–296. In *The Neuroscience of Morality: Emotion, Brain Disorders, and Development*. W. Sinnott-Armstrong, ed. Cambridge, MA: MIT Press.

McIntosh, D. N., Reichmann-Decker, A., Winkielman, P., and Wilbarger, J. L. (2006), When the Social Mirror Breaks: Deficits in Automatic, but Not Voluntary, Mimicry of Emotional Facial Expressions in Autism, *Developmental Science*, 9:295–302.

Morrison, D. and Gilbert, P. (2001), Social Rank, Shame and Anger in Primary and Secondary Psychopaths, *The Journal of Forensic Psychiatry*, 12:330–356.

Nucci, L. and Turiel, E. (1993), God's Word, Religious Rules, and Their Relation to Christian and Jewish Children's Concepts of Morality, *Child Development*, 64:1475–1491.

Ortony, A. and Turner, T. (1990), What's Basic About Basic Emotions?, *Psychological Review*, 97:315–331.

Phan, K. L. Wager, T. Taylor, S. F., and Liberzona, I. (2002), Functional Neuroanatomy of Emotion: A Meta-Analysis of Emotion Activation Studies in PET and fMRI, *NeuroImage*, 16: 331–348.

Ramirez, E. (2016), Neurosurgery for Psychopaths? The Problems of Empathy and Neurodiversity, *American Journal of Bioethics: Neuroscience*, 7:166–168.

Reisenzein, R., Studtmann, M., and Horstmann, G. (2013), Coherence between Emotion and Facial Expression: Evidence from Laboratory Experiments, *Emotion Review*, 5:16–23.

Russell, J. A. (1995), Facial Expressions of Emotion: What Lies beyond Minimal Universality?, *Psychological Bulletin*, 118:379–391.

Russell, J. A. (2016), A Skeptical Look at Faces as Emotion Signals, pp. 157–172. In *The Expression of Emotion: Philosophical, Psychological, and Legal Perspectives*. C. Abell and J. Smith, eds. New York: Cambridge University Press.

Turiel, E. (1977), Distinct Conceptual and Developmental Domains: Social Convention and Morality, *Nebraska Symposium on Motivation*, 25:77–116.

Wager, T. D., Kang, J., Johnson, T. D., Nichols, T. E., Satpute, A. B., and Feldman Barrett, L. (2015), A Bayesian Model of Category-Specific Emotional Brain Responses, *PLoS Computational Biology*, 11:e1004066.

Wong, Y. and Tsai, J. (2007), Cultural Models of Shame and Guilt, pp. 210–223. In *Handbook of Self-Conscious Emotions*. J. Tracy, R. Robins and J. Tangney, eds. New York: Guilford Press.

6

Anorexia Nervosa and the Embodied Mind

Lana Kühle

Introduction

Traditionally, philosophers of mind have been guided by a brainbound approach: the mind, whatever it turns out to be, will be related to or identical with the brain. The body, under this approach, plays a merely instrumental role—it is what keeps the brain alive and healthy. Over the past few decades there has been increasing resistance to the brainbound approach, and a strongly supported push for taking a non-brainbound approach: the body is not merely instrumental, but in many ways constitutive of the mind—the mind, whatever it turns out to be, is bodily in many fundamental respects. Thus, in explaining the mind, we must consider not only brain states, but also wider bodily processes. What those wider bodily processes are and how they constitute certain mental states are the key questions that we now face as non-brainbound—embodied mind—philosophers of mind.

Perhaps it is time to make a similar shift in psychiatry. There are certain disorders, such as anorexia nervosa, that have proven quite problematic to treat. The fifth edition of the *Diagnostic and the Statistical Manual of Mental Disorders (DSM-5)* classifies anorexia nervosa as an eating disorder—one consisting of behavioral and cognitive disturbances (APA 2013). The suggested treatment protocols have been focused on correcting those behavioral and cognitive problems. Yet, anorexia nervosa continues to have a very high relapse rate and a very high mortality rate. In short, we haven't been successful in treating this disorder. Perhaps the reason for our lack of success stems from the starting assumptions we've used to guide our understanding and treatments. Is anorexia nervosa properly classified as an eating disorder? Are its main disturbances mainly behavioral and cognitive? I propose here that we consider anorexia nervosa from an embodied mind approach.

If the mind is embodied, then we must look at how mental disorders might be bodily disorders. This has not been the standard approach in psychiatry. In what follows, I look at how anorexia nervosa has been understood and treated, and I present the data on how successful the standard practices have been to date. I then diagnose what I see as the cause of our lack of success and consider how an embodied mind approach can give us different, and arguably better, insight into the disorder. I look at evidence showing how this approach can deepen our understanding of the full spectrum of disturbances involved in the disorder and thereby help build treatment protocols that should prove more successful. Lastly, I return to philosophy and consider the cross-disciplinary benefits that come with a shift to an embodied mind paradigm in investigations of the mind.

Anorexia nervosa

The first description of anorexia nervosa (hereafter, AN) in the medical literature was given by Gull in 1874 (Gull 1874: 2974). It is classified as an eating disorder, and its most obvious symptom is an extremely low body weight. The National Institute of Mental Health (NIMH) characterizes AN as involving: extremely restricted eating, extreme thinness, a relentless pursuit of thinness and unwillingness to maintain normal/healthy weight, intense fear of gaining weight, distorted body image, self-esteem tied to perceptions of body weight and size, and denial of seriousness of low body weight (NIMH 2016). The *DSM*-5, similarly, characterizes AN along the following three main diagnostic features: persistent restriction of energy intake (which leads to extremely low body weight), intense fear of weight gain or becoming fat, and disturbance in self-perception of weight or shape (which causes perceptual distortions) (APA 2013: 339–340). As Keizer and colleagues show, "AN patients show significantly higher percentages of overestimation of their body size than [healthy controls]. […] Percentage of body size overestimation correlated with disease duration" (Keizer et al. 2013: 4). Body size and weight—the representation of, behaviors in relation to, and thoughts about—are taken to be the pivotal issues in the disorder.

The NIMH estimates that AN afflicts 0.6 percent[1] of the US adult population—between 2 and 4 million.[2] AN is a disorder that afflicts both men and women, although it is 2 ½ times more prevalent in women. Its onset typically occurs during teenage or young adult years—with incidence rates being their highest in females between the ages of 15 and 19 (Smink et al. 2012: 407)—and rarely occurs after the age of 40. AN has the highest mortality rate among all mental

disorders (ranging between 10 and 20 percent [Stinson 2017: 2]), with 1 in 5 among those that die doing so by suicide (Smink et al. 2012: 411). AN is notoriously difficulty to treat, with a response rate of approximately 50 percent even for the most effective treatments (Wilson et al. 2007), and for those who do respond to treatment there is a high relapse rate (approximately 30 percent [Carter et al. 2004]).

The first line of treatment typically focuses on the restoration of a healthy body weight. The American Psychiatric Association (APA)'s *Practice Guidelines for the Treatment of Patients with Eating Disorders* distinguishes three areas of treatment: nutritional rehabilitation, psychosocial interventions, and medication. In the APA's associated *Quick Reference Guide,* the following treatment goals are listed:

1. Restore or maintain healthy weight
2. Minimize food restrictions
3. Reduce binge eating and purging behaviors, if present
4. Provide education regarding healthy nutrition and eating patterns
5. Encourage healthy but not excessive exercise
6. Enhance the patient's motivation to cooperate and participate in treatment
7. Address underlying themes and correct core maladaptive thoughts and attitudes
8. Treat physical complications
9. Treat associated psychiatric conditions, including defects in mood regulation, self-esteem, and behavior
10. Provide family counseling and therapy where appropriate
11. Prevent relapse (APA 2006: 237–238).

Nutritional rehabilitation revolves around feeding therapies to address the nutritional and body weight symptoms, psychosocial treatments typically turn to cognitive behavioral therapy (CBT), and medication is used to correct weight. Although much of the initial treatment protocols and goals focus on nutritional and dietary changes, as is made clear above, it is believed that body image disturbances are what lie at the heart of the disorder.

The main line of treatment for body image issues has been CBT, which involves three aspects: cognitive, behavioral, and affective. The two most widely used CBT therapies are those developed by Rosen (1997) and Cash (1995). Both approaches "address three key aspects of body image disturbance: size perception (overestimation of overall size or size of a particular part, exaggeration of the extent of a perceived defect); cognitive and affective (negative, unrealistic, and

overvalued thoughts and emotions about the body, and the consequences of perceived defects); and behavioral (avoidance of anxiety provoking situations, checking and grooming behaviors, reassurance seeking)" (Farrell et al. 2006: 291). The overall goal of CBT is to help a person learn how to identify distorted or unhelpful thinking patterns, and recognize and change inaccurate beliefs. Insofar as AN is understood to arise from exaggerated or inaccurate beliefs about body size, thinness, and so on, CBT has been seen as the best approach to helping anorectics correct their disturbed beliefs.

As we can easily ascertain from the descriptions of the disorder and the treatment goals and protocols listed above, AN is understood as a disorder of eating behaviors and thoughts about eating and body size. In short, AN is seen as a behavioral and cognitive problem and has been treated as such. Yet, as the statistics also show, the outcomes of treatment are not satisfactory, with high relapse rates and a high mortality rate. Perhaps the reason for this is that we are taking the wrong approach and are misunderstanding the underlying issues in the disorder.

The problem and possible solution

Clearly we must do better—we need greatly reduced relapse and mortality rates. But, why have we not been able to do so? One reason may be that many anorectics don't seek help. Yet, among those that do seek treatment, and have done so on multiple occasions due to relapse, there must be a reason why success is so difficult to achieve. I propose that it has to do with the underlying assumptions at play in our understanding of the disorder. To date, the focus has been to change eating behaviors and correct cognitive disturbances related to body image. The assumption underlying this approach is that the root of the problem lies at behavioral (eating practices) and cognitive (beliefs about body size/weight) levels. However, there is reason to suspect that the behavioral and cognitive disturbances are merely consequences of further problems that themselves lie at the root of the disorder. If so, we must consider the possibility that our assumption is incorrect.

There is a lesson to be learned here by turning to philosophy. Philosophers of mind have notoriously been obsessed with the brain. At virtually every turn, contemporary philosophers of mind have taken a brainbound approach: whatever the nature of the mind turns out to be, it will be associated in some way with the brain. This view has been the driving hypothesis behind most,

if not all, scientific research on the mind in neuroscience, psychology, and cognitive science as well. When we think about the mind, we typically think of the brain. But is this the right approach? In the mid-1990s a few scholars proposed an entirely different way of looking at the relationship between the mind and the brain/body.[3] Influenced by Asian philosophical thought[4] as well as by work stemming from the Phenomenological tradition,[5] they argued that a brainbound approach is misguided and cannot properly account for the nature of the mind. The brain cannot be divorced from the body; rather the brain and body form an organic whole. If the mind is associated with our physical being, then it must be associated with the whole of it and not just a part of it. You cannot fully account for the mind, for our mental life, without considering the body. In short, we are embodied conscious organisms and must be taken as such when studying the mind in relation to the brain/body. The non-brainbound approach they proposed is now most commonly referred to as the embodied mind approach.[6] The mind, whatever it turns out to be, needs to be understood in terms of bodily processes, not merely brain processes. The embodied mind theory not only shifted the long-standing conversation about the mind and the brain because it pushed against strong intuitions, but it also allowed thinkers to make progress on certain problems and make sense of certain phenomena that brainbound theorists struggled with.[7] In particular, there has been promising work done in understanding the subjectivity of experience and the sense of self in bodily terms (Cassam 2011; Kühle 2014; 2017; Legrand 2006). Intuitively, we take our "self" to be fundamentally constituted by our minds—our thoughts, beliefs, desires, and so on. If we were to lose these mental states, or lose access to them—e.g., as in the case of severe Alzheimer's—we begin to think that the "self" is lost as well. These intuitions also underlie the religious beliefs of many whereby once the body dies the soul, or self, lives on. Yet, these are simply intuitions and if we explore the phenomenon of the self further, and are convinced by the benefits of taking on an embodied mind approach, then our views must change. Consciousness, and the self that is associated with it, cannot be relegated to purely mental states. The self will turn out to be bodily just as consciousness is. Looking back to the example of Alzheimer's, indeed the person afflicted by this disease has lost elements of herself in losing access to much of her mental states, but her self remains insofar as she remains a bodily subject of experience.

As has been the case in philosophy, psychiatry has also long been caught in the paradigmatic view of the mind as brainbound. Moreover, psychiatry is a cognitive and behaviorally based discipline. To be sure, it does consider the biological

factors at play in mental disorders (virus, genetics, etc.), but classification and treatment of disorders are often focused on changing aberrant behaviors and thoughts. Mental disorders have to do with disrupted cognitive systems, and cognitive systems are seated in the brain. This paradigmatic approach may be the problem—AN may not be best understood as an essentially behavioral and cognitive disturbance. Indeed it may be time to shift the paradigm in psychiatry in the same way that many have begun to do in philosophy. There is strong reason to think that AN would be better understood as an embodied disorder, involving disturbances at the level of body perceptions. To clarify, body perception is not the same as body image—a concept that has been often referred to in the AN literature in psychiatry. A body image is a representation, "a mental image of one's body that could be captured in an objective way, almost as a 'picture'" (Gaete and Fuchs 2016: 19). However, our awareness of our body is not reducible to such a representation. There is a difference between the mental image one forms about one's body—a belief or thought—and how one *experiences*, or *feels* one's body. AN appears to involve not only the disruption of one's body image, but there is reason to suspect that the disturbed body image is the result of a disturbance in how one experiences one's body—a disturbance in body perception. In other words, the anorectic doesn't only incorrectly believe her body to be of a certain size—far larger than it actually is—she also cannot feel her body properly—she doesn't feel her inner, visceral body, for example heartbeat, hunger, and so on, in the way that a non-anorectic does. The anorectic feels a distance between herself and her body in a way that non-anorectics don't—a reduced body perception.

The suggestion that the issue involves body perception, and requires a shift away from considering AN to be primarily a body image disturbance, is supported by the subjective experience of anorectics. Their descriptions clearly reflect that the main experiential factor is one involving how they *feel*, how their body creates a disturbed *sense of self*. Their sense of self as bodily is disrupted. They feel themselves as being a certain way, and this leads them to hold certain beliefs about how they ought to be. Their experience is, in effect, a disturbance of embodiment. I consider here that AN is not solely, or even mainly, a disturbance in one's *beliefs* about body size and eating behaviors, but rather a disturbance in one's *bodily awareness* at non-cognitive levels. It is an embodiment disorder. To further motivate this, let's look closer at the anorectic's experience.

Anorectics often feel misunderstood. What they're feeling, experiencing, their intentions, and so on do not line-up with how others describe the problem. The missed connection between the patient's experience and the objective understanding of the disorder only serves to further isolate the anorectic, and

may push her away from treatment rather than toward it. If one reads anorectics' reports about what it's like to have AN, one quickly notices that it's not about eating. It is abundantly clear that, as Stinson explains, "[t]he people making these testimonials feel misunderstood by their families, friends and doctors. One of the women diagnosed with AN interviewed in Malson (1998) claims that 'a lot of psychiatrists don't know what they're talking about … it shouldn't be looked at on the surface. It shouldn't be looked at as an eating, just an eating problem.'" (Stinson 2017: 3). Rather, anorectics often describe their experience as involving non-cognitive aspects. Patients in Charland and colleagues' *anorexia experience study*, gave the following descriptions of their experiences:

> That's another thing that's—I mean I wouldn't say anorexia is a thought as such. I don't think I, it's more of a feeling I, I WANT to do it and I guess in a way it's almost an EMOTION, anorexia. (P16)[8]
>
> I can still see facts as they are. What I've lost is the ability to apply them for myself because emotions have taken over, but I don't think my rational side is being destroyed, or anything. I think it's still there, it's just a question of accessing it. (P17)
>
> If I saw like a bin with an empty chocolate bar wrapper I would start shaking, I would start crying, I would be just in complete and utter torment and that's what it does to you. (P13) (Charland, Hope, and Tan 2013: 357–358)

These are but a few among many descriptions that reveal a clear dissociation between the anorectic's experience of the disorder, and the manner in which the disorder is being characterized. There is a disconnect between the subjective, experiential side of AN and the objective, scientific conceptualization of AN as a cognitive disorder. Such a dissociation can explain the lack of success at treating the disorder. If we analyze the experiential reports we clearly see a reason for this. What appears to be of central issue is, at least in part, an emotional disturbance which leads to a cognitive disturbance. Consider Charland and colleagues' discussion of the last quoted report above:

> [i]ndeed, the person knew, intellectually, that she could not gain weight simply by walking close to a chocolate wrapper. Rather the person primarily experiences anxiety, even terror. If probed the person might then relate this to beliefs about weight gain but *the primary experience is of fear*. Indeed, the beliefs may be almost like rationalizations of the fear rather than its cause. (Charland et al. 2013: 362; emphasis added)

To be sure, these are but a select few among many reports[9]; nevertheless, there is insight to be gained. First, the subjective experience of the anorectic has been

largely neglected, and this has led to an inability to correctly understand the disorder. Second, if we take the anorectics' subjective experience of AN as our starting point, we quickly see that our approach needs to change. AN is far more than a disturbance in one's beliefs about weight and thinness, or about one's image of one's body. AN, at its core, may not be a cognitive issue at all. If that is the case, then treatments aimed at correcting false beliefs and other cognitive problems are sure not to be successful: "Providing information about healthy body weights, or attempting to change their minds about beauty ideals is largely ineffective, because the problem is not as simple as holding false beliefs about healthy eating" (Stinson 2017: 3). Yet, if AN is not mainly a cognitive issue, then what sort of issue might it be? As I've briefly suggested, and will now explore in greater detail, AN is best understood as an embodiment issue.

The embodied mind view holds that the mind is constituted by the brain and wider bodily processes. As a result, a disorder of the mind can be understood as a disorder of the brain/body—in short, a disorder of embodiment. Looking at AN as a disorder of embodiment makes far better sense of the anorectics' subjective experience. The disorder has to do with how one perceives one's body in relation to one's sense of self—bodily self-awareness. Working with this view, we can look for disturbances having to do with bodily self-awareness in AN.

Bodily self-awareness stems from a variety of sensory modalities across the body. One can see or touch one's body and become aware of it. One can also feel one's body as being seated or having one's legs crossed without having to look or touch the body. This sense of the body's position in space—and whether the body is moving and our sense of balance—are attributable to proprioception. One can also feel one's heartbeat, a full bladder, a bloated stomach, or being hot. This sense of the body's inner, visceral environment is attributable to interoception. Some[10] have also argued that the feeling of an emotion—e.g., the racing heart, the flush of warmth across the face, the tightness in the chest, the lump in one's throat, and so on of panic—is attributable to interoception.[11] Typically, in our day-to-day activities our sight, touch, hearing, smell, and taste are all involved with the environment outside our bodies. That is to say those sensory modalities are world directed, not body directed. Proprioception and interoception, however, always remain body directed, and are thus the main sources of bodily self-awareness. As such, if we want to understand where the disturbances in body perception lie in AN, we may want to consider those sensory modalities.

Many have done just that, and have shown the important link between body-oriented sensory modalities and AN. Charland and colleagues (2013) propose that the core issue in AN might have to do with anxiety—which, as an emotion,

has to do with interoception. Herbert and Pollatos and colleagues (2012) discuss the possibility that AN might have to do with a deficit in interoceptive awareness, with anorectics having a reduced implicit awareness of their inner body. Keizer and colleagues (2013) propose that the issue may reside in tactile and somatosensory problems associated with body perception. Case and colleagues (2012) suggest that the distorted body perceptions may have to do with problems at the level of multisensory integration. In short, it's not clear where the heart of the issue lies, however what is clear from all these proposals is that we need to consider AN as an embodied disorder, rather than as a mental disorder.

One very promising line of inquiry lies with interoception and its relation to bodily self-awareness. Our sense of embodiment—of being a living, flesh and blood organism—is driven by interoception. Interoception is linked to bodily self-awareness in two crucial ways: via emotion experiences and via our sense of self as embodied. In regard to the experience of emotions, as mentioned, we are said to feel an emotion by feeling the various interoceptive bodily changes characteristic of that emotion (Bennett and Hacker 2005). The feeling of fear just is the feeling of an increase in heart rate, muscle tension, shortening of breath, the rush of adrenaline, and so on. Interoception is also linked to our sense of self. There is ever-increasing evidence showing a strong correlation between specific areas of the brain that consolidate the interoceptive bodily signals into a coherent representation of the current state of the body as a whole and one's sense of self. Disruptions in these areas—disruptions in interoceptive awareness—correlate to disruptions in the sense of self (Herbert et al. 2012). Problems with interoceptive awareness can be manifested in a number of ways, one of which being that one can become less attuned to their inner body and thereby have a reduced interoceptive sensitivity. This manifests as a muted experience of sensory stimulation from certain interoceptive receptors, or as a weakened connection to one's body such that one can be more easily led to experience non-bodily objects as one's own body—as exemplified by the rubber hand illusion (RHI).[12]

With the crucial role played by interoception in two areas that are of concern in AN—emotions and body awareness—it seems a likely avenue of investigation as a source of the disturbances. We would expect to find that interoception, and more specifically, one's awareness of one's inner, visceral body—interoceptive awareness—is disturbed in AN. If this is correct then treating the disturbances in bodily processes associated with AN may prove more successful in treating the disorder. As I shall now turn to consider, there is growing research that takes an embodied mind approach to AN, and the findings that result from it are striking.

Embodied Mind Theory and AN—A new understanding

As mentioned, interoceptive awareness is associated with our feeling of emotions and with our sense of bodily self. If we study interoceptive awareness in AN, along both of these aspects, we find interesting results. The *DSM*-5 lists anxiety disorder as one of the most common co-occurring disorders with AN. Subjective reports make clear that AN involves strong feelings of anxiety and other negative emotions. More, as can be clearly ascertained in anorectics' subjective reports, anxiety along with fear and panic are central elements of what AN feels like. The research also shows that anorectics have diminished interoceptive awareness as associated with the sense of self. Their sensitivity to their inner visceral body is reduced compared to healthy controls (Fassimo et al. 2004; Mastumoto et al. 2006; Pollatos et al. 2008). They are more susceptible to the RHI (Eshkevari et al. 2012), and susceptibility to the RHI is correlated with a reduced interoceptive sensitivity (Tsakiris et al. 2011). Their perception of affective touch is reduced (Crucianelli et al. 2016).Further, the area responsible for consolidating interoceptive signals into an overall interoceptive bodily self-awareness shows altered neural activity in AN (Kerr et al. 2016). In short, anorectics present with a heightened awareness and feeling of negative emotions, as well as a muted interoceptive awareness of the body. How can this be so? On the one hand, they seem to have an increased interoceptive awareness—in feeling of emotions—and on the other hand, they appear to have a reduced interoceptive awareness—in relation to their sense of bodily self. There is an interpretation of this seemingly contradictory data that makes clear sense of the underlying disturbances in the disorder. The suggestion is that deficits in interoceptive awareness are the result of the anorectic's inability to cope with negative emotions. Anxiety and panic are known to bring about a fight or flight response. In AN, the heightened feeling of anxiety and panic brings about a flight response. The anorectic responds to the heightened feelings of anxiety by creating a distance with her bodily feelings. If the feeling of anxiety is the feeling of the inner body, and one wants to remove oneself from these feelings, then one might do so by distancing oneself from one's inner body. As Gaete and Fuchs explain, "This creates a relieving distance from unpleasant experiences […]. The studies of Skårderud as well as those of Atwood and Stolorow support the idea of an objectification of the body as a respite from the negative feelings" (Gaete and Fuchs 2016: 22). In distancing herself from the interoceptive signals, she also reduces her overall interoceptive awareness. However, given the link between interoceptive awareness and the sense of self as embodied—i.e., one's perception of the body as one experiences it—the anorectic is then forced to

shift her attention to a more objective image of the body. This is what creates the detachment and brings forth the issues at the cognitive level in relation to the body image. As Bowden concludes from phenomenological study of AN, "[t]he anorectic appears to lose the tacit background understanding of her body, instead finding out about it through a disassociated, 'watching'. It seems that it has almost entirely taken upon the role of perceiv*ed* [object], losing the perceiv*ing* [subject] nature that is usual to the body" (Bowden 2012: 234). The move away from the body as it is *experienced* and toward the body as it is *conceived*—from the subjective body to the objective body—is a defense mechanism.

If this is the correct account of what happens in AN, then it is clear why focusing on the behavioral and cognitive components of the disorder won't be sufficient. The root of the issue is not cognitive. The disturbed cognitions are a result rather than the cause of a disturbed bodily self-awareness. The anorectic, in short, must reconnect with and learn to handle their embodiment. As one anorectic explained it: "It is strange, because during the whole day at the unit, I try to avoid feeling and relating to my body. While in the body awareness group, we are encouraged to be aware of the body and its reactions. I relax and can feel my body in a positive way. I can feel my breath is getting deeper, and I manage to let tension go, and this is a good feeling. At the same time, it's like I feel my body too strongly" (Kolnes 2012: 284). Importantly, in addressing the subject's disconnection with their lived body we must also consider the negative feelings that the anorectic was fleeing from, as restoring a healthy bodily self-awareness will also bring back to the fore those negative emotions. Treatment, then, must include helping the anorectic cope with negative feelings in addition to reconnecting with her feeling body.

Encouragingly, there is evidence showing that improvements in interoceptive awareness correspond to a reduction in symptoms of AN at long-term follow-up (Merwin et al. 2010). Unfortunately, this line of treatment remains under developed, under pursued, and under researched.

Working hand in hand

The value of shifting the paradigm in psychiatry as has been done by many in philosophy extends beyond understanding and treating AN. The embodied mind approach can also afford deeper insights into other disorders involving disruptions in body perception—e.g., somatoparaphrenia and body dysmorphic disorders. It can also help make sense of mental disorder such as schizophrenia

and severe depression (Maiese 2011, 2012, 2016). Importantly, the direction of influence does not only go one way. It is not only psychiatry that can benefit from making a shift to an embodied mind view. Philosophers of mind, especially those concerned with accounting for consciousness as embodied, will benefit from cross-disciplinary investigations into embodiment as well. As embodied mind thinkers, we face the following questions: Which are the wider bodily processes that constitute the mind and how do they constitute certain mental states? One avenue of investigation that many have taken to answer these questions has been to look at bodily awareness and its relation to our sense of self, and to how our body and our awareness of our body more specifically constitute our sense of self and our subjective experience of the world. A clear way of investigating these issues is to consider pathological cases of body awareness, cases such as AN. "Progress of brain research during the past two decades demonstrates the power of the neurobiological paradigm. When applied to the phenomenon of mental illness, however, this progress often generates a restricted perspective, as is typical for any scientific paradigm" (Fuchs 2009: 220). In the end, it is important to keep the lines of cross-disciplinary discussion open, and to prevent becoming pigeonholed in our views. The embodied mind approach may be what's needed to take the next steps in psychiatry. As Kolnes so nicely puts it: "Embodying the body in patients with anorexia nervosa [and other mental disorders] may help stabilize the body *and* the mind, and to discover unknown resources within the patients. Body awareness and body contact needs to first be established for the patients to be able to attune themselves to their emotions and minds, as well as to the surrounding world. Clearly, the body deserves more attention in treatment programs for eating disorders [and other mental disorders]" (2012: 287).

Acknowledgments

I would like to thank Şerife Tekin and Robyn Bluhm for their very helpful comments and support.

Notes

1 Estimates based on the definition in the latest *DSM*-5 range from 2 to 4.3 percent prevalence among women. See Stinson (2017: 2)

2 Estimates of the disorder prevalence are difficult to make as many who suffer from AN don't seek treatment. NIMH estimates that only 33.8 percent of anorectics are

receiving treatment. This may be due to the continued societal view that thinness is a valued aesthetic feature, and thereby isn't seen as wrong or problematic. If we combine this view with the evidence that anorectics have a disturbed perception of their actual body size, then many don't know or refuse to acknowledge that there is an issue with their body size and behaviors or thoughts regarding it.

3 The seminal text that brought forward the shift in the discussion is Varela, Thompson, and Rosch (1991).

4 In particular, Buddhist philosophy.

5 The Phenomenological Tradition refers to a school of thought associated with thinkers such as Edmund Husserl, Maurice Merleau-Ponty, Jean-Paul Sartre, Martin Heidegger, and so on.

6 Since it was first proposed, there have been a number of different non-brainbound views that have appeared—embodied, enacted, extended, embedded. These differ in important ways and are not all compatible with one another. It is beyond the scope of this chapter to discuss these differences and compatibilities, however I refer interested readers to see Mark Rowlands (2010) for further discussion. For my purposes, I shall refer to the embodied mind view in a broad sense as reflecting the shift to a non-brainbound view.

7 For example, in the case of explaining discrepancies in visual development as shown in Held and Hein's kitten experiment (Held and Hein 1963) or in accounting for the adaptation in visual perception when wearing inverted goggles.

8 "P" followed by a number was used to identify subjects in the study.

9 See Charland et al. (2013), Stinson (2017) and Malson (1998) for further discussion of anorectics' reports of their experiences.

10 This view dates back to James (1884) and Lange (1885). More contemporary thinkers that defend this view, or a version thereof, are Antonio Damasio (1999), Craig (2009), Kühle (2014, 2017).

11 The act of running, for example, would be the behavioral component of panic, and the conceptualization of what one is feeling as "panic" would be the cognitive component.

12 The RHI is a perceptual illusion that arises when visual information dominates over tactile and proprioceptive information in a multisensory experience of seeing a rubber hand being stroked while having your hand stroked simultaneously. For some, vision dominates over the multisensory input causing the illusion to come about wherein one feels that the rubber hand IS one's hand. See Botvinick and Cohen (1998) for the original study of the RHI.

References

American Psychiatric Association (2006), *Treating Eating Disorders—A Quick Reference Guide*, Arlington, VA: American Psychiatric Association Publishing.

American Psychiatric Association (2013), *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Arlington, VA: American Psychiatric Association Publishing.

Bennett, M. R. and Hacker, P. M. S. (2005), Emotion and Cortical-Subcortical Function: Conceptual Developments, *Progress in Neurobiology*, 75:29–52.

Botvinick, M. and Cohen, J. (1998), Rubber Hands 'Feel' Touch That Eyes See, *Nature*, 391:756.

Bowden, H. (2012), A Phenomenological Study of Anorexia Nervosa, *Philosophy, Psychiatry, & Psychology*, 19:227–241.

Carter, J. C., Blackmore, E., Sutandar-Pinnock, K., and Woodside, D. B. (2004), Relapse in Anorexia Nervosa: A Survival Analysis, *Psychological Medicine*, 34:671–679.

Case, L. K., Wilson, R. C., and Ramachandran, V. S. (2012), Diminished Size-Weight Illusion in Anorexia Nervosa: Evidence for Visuo-Proprioceptive Integration Deficit, *Experimental Brain Research*, 217:79–87.

Cash, T. F. (1995), *What Do You See When You Look in the Mirror? Helping Yourself to a Positive Body Image*, New York: Bantam Books.

Cassam, Q. (2011), The Embodied Self. In *Oxford Handbook of the Self*. Shaun Gallagher, ed. Oxford: Oxford University Press.

Charland, L. C., Hope, T., Stewart, A., and Tan, J. (2013), Anorexia Nervosa as a Passion, *Philosophy, Psychiatry, and Psychology*, 20:353–365.

Craig, A. D. (2009), How Do You Feel—Now? The Anterior Insult and Human Awareness, *Nature Reviews Neuroscience*, 10:59–70.

Crucianelli, L., Cardi, V., Treasure, J., Jenkinson, P. M., and Fotopoulou, A. (2016), The Perception of Affective Touch in Anorexia Nervosa, *Psychiatry Research*, 239:72–78.

Damasio, A. (1999), *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*, Orlando, FL: Harcourt Books.

Eshkevari, E., Rieger, E., Longo, M. R., Haggard, P., and Treasure, J. (2012), Increased Plasticity of the Bodily Self in Eating Disorders, *Psychological Medicine*, 42:819–828.

Farrell, C., Shafran, R., and Lee, M. (2006), Empirically Evaluated Treatments for Body Image Disturbance: A Review, *European Eating Disorders Review*, 14:289–300.

Fassimo, S., Pierò, A., Gramaglia, C., and Abbate-Daga, G. (2004), Clinical, Psychopathological and Personality Correlates of Interoceptive Awareness in Anorexia Nervosa, Bulimia Nervosa and Obesity, *Psychopathology*, 37:168–174.

Fuchs, T. (2009), Embodied Cognitive Neuroscience and Its Consequences for Psychiatry, *Poiesis and Praxis*, 6:219–233.

Gaete, M. I. and Fuchs, T. (2016), From Body Image to Emotional Bodily Experience in Eating Disorders, *Journal of Phenomenological Psychology*, 47:17–40.

Gull, W. W. (1874), Anorexia Nervosa (Apepsia Hysterica, Anorexia Hysterica), *Transactions of the Clinical Society of London*, 7:22–28.

Held, R. and Hein, A. (1963), Movement-Produced Stimulation in the Development of Visually Guided Behavior, *Journal of Comparative and Physiological Psychology*, 56:872–876.

Herbert, B. M. and Pollatos, O. (2012), The Body in the Mind: On the Relationship between Interoception and Embodiment, *Topics in Cognitive Science*, 4:692–704.

James, W. (1884), What Is an Emotion?, *Mind*, 9:188–205.

Keizer, A., Smeets, M. A. M., Dijkerman, H. C., Uzunbajakau, S. A., Van Elburg, A., and Postma, A. (2013), Too Fat to Fit through the Door: First Evidence for Disturbed Body-Scaled Action in Anorexia Nervosa during Locomotion, *PLOS One*, 8:1–7.

Kerr, K. L., Moseman, S. E., Avery, J. A., Bodurka, J., Zucker, N. L., and Simmons, W. K. (2016), Altered Insula Activity during Visceral Interoception in Weight-Restored Patients with Anorexia Nervosa, *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, 41:521–528.

Kolnes, L. J. (2012), Embodying the Body in Anorexia Nervosa—A Physiotherapeutic Approach, *Journal of Bodywork and Movement Therapies*, 16:281–288.

Kühle, L. (2014), *Embodiment and Subjectivity—The Origins of Bodily Self-Awareness*, PhD dissertation, University of Toronto.

Kühle, L. (2017), The Subjectivity of Experiential Consciousness: It's Real and It's Bodily, *Mind and Matter*, 15:91–109.

Lange, C. G. (1885), The Emotions: A Psychophysiological Study. In *The Emotions*. I. A. Haupt, trans. K. Dunlap, ed. Reprinted (1992). Baltimore, MD: Williams & Wilkins Company.

Legrand, D. (2006), The Bodily Self: The Sensorimotor Roots of Pre-Reflective Self-Consciousness, *Phenomenology and the Cognitive Sciences*, 5:89–118.

Maiese, M. (2011), *Embodiment, Emotion, and Cognition*, London: Palgrave Macmillan.

Maiese, M. (2012), Rethinking Attention Deficit Hyperactivity Disorder, *Philosophical Psychology*, 25:893–916.

Maiese, M. (2016), *Embodied Selves and Divided Minds*, Oxford: Oxford University Press.

Malson, H. (1998), *The Thin Woman: Feminism, Post-Structuralism and the Social Psychology of Anorexia Nervosa*, East Sussex, NY: Routledge.

Matsumoto, R., Kitabayashi, Y., Narumoto, J., Wada, Y., Okamoto, A., Ushijima, Y., Yokoyama, C., Yamashita, T., Takahashi, H., Yasuno, F., and Suhara, T. (2006), Regional Cerebral Blood Flow Changes Associated with Interoceptive Awareness in the Recovery Process of Anorexia Nervosa, *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 30:1265–1270.

Merwin, R. M., Zucker, N. L., Lacy, J. L., and Elliott, C. A. (2010), Interoceptive Awareness in Eating Disorders: Distinguishing Lack of Clarity from Non-Acceptance of Internal Experience, *Cognition and Emotion*, 24:892–902.

National Institutes of Mental Health (2016), Eating Disorders, revised February 2016, https://www.nimh.nih.gov/health/topics/eating-disorders/index.shtml, section 2—'Signs and Symptoms, Anorexia Nervosa'.

Pollatos, O., Kurz, A. L., Albrecht, J., Schreder, T., Kleeman, A. M., Schöpf, V., Kopietz, R., Wiesmann, M., and Schandry, R. (2008), Reduced Perception of Bodily Signals in Anorexia Nervosa, *Eating Behaviors*, 9:381–388.

Rosen, J. C. (1997), Cognitive-Behavioral Body Image Therapy. In *Handbook of Treatment for Eating Disorders*, 2nd edition. D. M. Garner and P. E. Garfinkel, eds. New York: The Guilford Press.

Rowlands, M. (2010), *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*, Cambridge, MA: MIT Press.

Smink, F. R. E., Van Hoeken, D., and Hock, H. W. (2012), Epidemiology of Eating Disorders: Incidence, Prevalence and Mortality Rates, *Current Psychiatry Reports*, 14:406–414.

Stinson, C. (2017), The Absent Body in Psychiatric Diagnosis, Treatment, and Research, *Synthese*:1–24, https://link.springer.com/article/10.1007/s11229-017-1507-8.

Tsakiris, M., Tajadura-Jimenez, A., and Costantini, M. (2011), Just a Heartbeat Away from One's Body: Interoceptive Sensitivity Predicts Malleability of Body-Representations, *Proceedings of the Royal Society B: Biological Science*, 278:2470–2476.

Varela, F. J., Thompson, E., and Rosch, E. (1991), *The Embodied Mind: Cognitive Science and Human Experience*, Cambridge: MIT Press.

Wilson, G. T., Grilo, C. M., and Vitousek, K. M. (2007), Psychological Treatment of Eating Disorders, *American Psychologist*, 62:199–216.

Study Questions

1. *Are mental disorders brain disorders?*
2. *Does mental illness change the identity of the person over time?*
3. *Do people with mental disorders have free will?*
4. *Do people with mental disorders lack rationality? Under what circumstances?*
5. *What are the implications of multiple realizability for mind-brain identity theory?*

Part Three

Phenomenology and Psychiatry

7

Merleau-Ponty and the Foundations of Psychopathology

Anthony Vincent Fernandez

Maurice Merleau-Ponty was a twentieth-century French philosopher who worked at the intersection of phenomenology and existentialism. Phenomenology, founded by Edmund Husserl and further developed by his students, including Martin Heidegger, is concerned with human experience and existence. Traditionally, it describes the essential structures of consciousness—i.e., those features that hold for any experiencing human subject—including selfhood, intersubjectivity, affectivity, and temporality. Existentialism, in contrast with phenomenology, is not a systematic research program. Its themes originate in the nineteenth-century work of Søren Kierkegaard and Friedrich Nietzsche, though "existentialism" was not used as a philosophical label until Gabriel Marcel applied it to the work of Jean-Paul Sartre in the 1940s (Fulton 1999: 12–13). Like phenomenology, existentialism is the study of the nature of human existence and with how we experience a meaningful world. But existentialists are primarily concerned with human processes of self-creation—such as how we become who we are, or transform our identity—rather than with the unchanging *essence* of human being. In this respect, existentialists typically study contingent, rather than necessary, features of human existence.

In integrating these two apparently opposing lines of thought, Merleau-Ponty belongs in the company of other French phenomenologists, including Jean-Paul Sartre and Simone de Beauvoir, who were inspired by Heidegger's application of phenomenology to the study of existential themes, such as death and authenticity. These philosophers continued phenomenology's inquiry into the structures of experience, but turned their attention toward contingent and particular structures, rather than just necessary and universal ones. This existential turn in phenomenology is perhaps best expressed by Sartre's famous line: "existence precedes essence" (Sartre [1945] 2007: 20). According to

Sartre, the central feature of human existence is that we are free to decide who we are—we are radically free. I might see myself as a waiter, a mechanic, or a professor but, in truth, I am none of these. I am, first and foremost, the power of freely willing myself to become who I wish to be. (In an interview, Beauvoir claimed that Sartre even believed seasickness was the product of weak will; with enough willpower, anyone could overcome his seasickness [Simons 1992: 30].) Merleau-Ponty, a friend and colleague of Sartre and Beauvoir, also made human contingency central to his philosophical project—for him, existence also precedes essence. But Merleau-Ponty was more concerned with how our natural and cultural circumstances shape us, than with how we freely shape ourselves.[1] To this end, he was especially interested in psychopathological conditions, from psychoanalytic neuroses to neurological disorders.

Merleau-Ponty's studies of psychopathology are often read as nothing more than instrumental examples in his studies of perception, including in his major work, *Phenomenology of Perception*; by examining how some aspect of perceptual experience goes awry (for instance, in schizophrenic hallucinations), we can gain insight into aspects of normal perception that we might have otherwise overlooked. This interpretation is not without merit (Merleau-Ponty does employ psychopathological examples instrumentally), but it would be inaccurate to characterize his interest in psychopathology as *merely* instrumental. Merleau-Ponty was interested in the psychopathological forms of human experience and existence because he aimed to construct a phenomenological program that adequately accommodated these conditions.

There are, of course, many phenomenologists who devoted themselves to the study of psychopathological conditions. Karl Jaspers first established psychopathology as an interdisciplinary program employing the tools of natural science alongside philosophical phenomenology. In *General Psychopathology*, he draws a distinction between "explanation" and "understanding," arguing that scientists can *explain* the causes of disorder, but only phenomenologists can *understand* the experience of disorder (Jaspers [1913] 1997). Only through the integration of natural science and phenomenology can we give an adequate account of mental disorders. Numerous phenomenological psychopathologists followed this line of thinking, including Ludwig Binswanger, Medard Boss, Frantz Fanon, Kimura Bin, Eugene Minkowski, Erwin Straus, and Hubertus Tellenbach. They developed novel accounts of mental disorders, including schizophrenia, melancholia, and various neuroses. With such a rich tradition of phenomenological psychopathology, why should we devote ourselves to the study of Merleau-Ponty's approach?

Merleau-Ponty's approach is distinctive because he seriously engages with the philosophical foundations of his own work. Most other phenomenological psychopathologists either took phenomenology's ability to illuminate psychopathological experience for granted, or they integrated phenomenological and empirical approaches without considering the philosophical implications of these integrations.[2] This is why we must turn to Merleau-Ponty, who explicitly developed metaphysical and methodological foundations for phenomenological psychopathology.

In this chapter, I outline some distinctive elements of Merleau-Ponty's metaphysics and method, and show how his work can provide a foundation for contemporary phenomenological studies of mental disorder. In the next section, I argue that Merleau-Ponty developed a new *metaphysics* that accommodates psychopathology by challenging the traditional distinction between transcendental and empirical phenomenology. In the subsequent section, I argue that Merleau-Ponty modified the core of the phenomenological *method*—i.e., the reduction, or epoché—to critically engage with the human sciences. Each section is composed of three subsections: the first provides an account of Husserl's phenomenology, and shows how it falls short of what we require for the study of psychopathology; the second describes Merleau-Ponty's novel phenomenological approach; and the third provides examples of contemporary applications of Merleau-Ponty's approach to psychopathology.

Metaphysics

It has been argued that Merleau-Ponty began his philosophical career as a phenomenologist, but by the end he had become a metaphysician—i.e., a philosopher whose primary concern is with the nature of being or reality. Merleau-Ponty certainly devoted himself to metaphysical questions in his later works, but he was concerned with metaphysics from the start. In this section, I outline Merleau-Ponty's early phenomenological metaphysics, and show how he developed it with the aim of doing justice to psychopathological forms of human experience and existence.

Husserl's Transcendental Idealism

Before jumping into Merleau-Ponty's metaphysics, it will be helpful to outline the metaphysical position that he aims to overcome. Rather than criticizing a

single philosopher, Merleau-Ponty criticizes a philosophical position called transcendental idealism. This metaphysical stance originated in Immanuel Kant's *Critique of Pure Reason* ([1781] 1999), and had a major influence on the course of nineteenth- and twentieth-century philosophy. While transcendental idealism takes various forms, we can focus on two elements that concern Merleau-Ponty specifically: First, it posits that the human subject constitutes or discloses a lived, orderly, and meaningful world—or makes the experiential world present—through some operation of the mind; second, it posits that the structures of human subjectivity that constitute the world of experience are essential, unchanging, or innate mental structures—they are necessary conditions for the possibility of knowledge and experience.

Husserl, insofar as he aimed to uncover the *essential* structures of consciousness and the necessary conditions for the possibility of experience, espoused a metaphysics of transcendental idealism. In this respect, phenomenology was originally conceived as an eidetic science—a philosophical investigation into the *eidos*, or essence, of human experience and existence. This approach is best exemplified in *Ideas I* (Husserl [1913] 2014), the first major work in which Husserl established phenomenology as a transcendental project inspired by the Kantian tradition. In this early work, Husserl was not concerned with human contingency, variability, or particularity. Instead, he concerned himself with necessity, invariance, and universality.

But soon his phenomenology became "genetic phenomenology," a concrete investigation into the genesis—i.e., the origins and development—of human experience. This new phenomenology emphasizes how our body and history shape how we experience and understand our environment. Developing phenomenology in this direction, Husserl explores physical disabilities, such as blindness and deafness, as well as cultural differences. He admits, for example, that some cultural groups have a cyclical—rather than linear—conception of time, and therefore experience the flow of time differently (Husserl [1935] 2008). But these studies introduced confusion into his philosophical project: How can we articulate the essential, invariant structure of experience if this structure differs across cultures and for people with certain physical disabilities? This seems a contradiction. He resolved this concern in his final book, *The Crisis of European Sciences and Transcendental Phenomenology* (Husserl [1954] 1970), by arguing that there are both essential, unchanging structures of experience and historically contingent structures. Necessary structures are "transcendental"; contingent structures are "empirical." Furthermore—and this is the key point in resolving the contradiction—the empirical structures

are founded upon the transcendental structures. That is to say, despite our differences, there are basic structures of experience that we all share. If, for example, you and I grew up in different linguistic communities, then I will perceive certain symbols etched on paper as meaningful words, but you will perceive them as meaningless symbols (if you perceive them as symbols at all). You, on the other hand, will experience other kinds of symbols as meaningful words, while I will experience them as just meaningless symbols. Yet, in spite of this difference, we both have the capacity to acquire written language; that is, to experience symbols etched on paper as conveying multifaceted and complex meanings. In other words, while the linguistic structure of our experience differs *empirically*—perhaps I understand English and you understand Mandarin—we both have the same *transcendental* conditions that allow us to acquire and employ written language. In the same spirit, we could resolve the apparent conflict between different cultures' experiences of the passage of time by suggesting that empirically different conceptions (e.g., of time as linear or as circular) are founded upon, or made possible by, the same transcendental structure of temporality.

This two-level approach seems to accommodate human difference, thereby allowing phenomenologists to investigate diverse forms of human life. But Husserl was not convinced that his solution could accommodate more profound changes in experience, such as the experience of the infant, or of the person with a severe mental disorder (Husserl [1954] 1970: 187–188). These conditions might involve alterations of the most fundamental structures of experience. The infant, for example, lacks the capacity for written language—not because she has not learned one, but because she currently lacks the cognitive conditions for learning a written language in the first place (Heinämaa 2014). If these alterations occur not only in the course of normal development, but also in psychopathological conditions in adulthood, then this might undermine Husserl's claim that there are universal structures of experience shared by all human beings.

Merleau-Ponty's Metaphysics

Husserl does not resolve this tension in his own work, but he sets it out as a problem for future phenomenologists. Luckily, his challenge was not ignored. Merleau-Ponty was one of the first philosophers to acquaint himself with Husserl's late work and unpublished manuscripts, and took up Husserl's abovementioned challenge. He resolved to develop a new metaphysics that accommodated the diversity of human life, including psychopathology.

In Merleau-Ponty's most famous work, *Phenomenology of Perception* ([1945] 2012), he introduces the case of Johann Schneider, a German First World War veteran who suffered a brain injury and whose condition was studied extensively by the psychologist Adhemar Gelb and the neurologist Kurt Goldstein. Schneider's condition involves a fundamental disturbance in his bodily motility. He is unable to perform "abstract" movements with his eyes closed—i.e., movements that take place in an imagined or merely possible situation, rather than the actual, concrete situation. Such actions might include anything from moving a limb up and down upon command to performing in a theatrical production to illustrating how objects move in an objective coordinate system. Yet, in spite of Schneider's inability to perform these movements, he can easily perform "concrete" movements—i.e., movements that take place within his actual, or concrete, situation. While Schneider cannot move his arm up and down upon command, he can easily reach into his back pocket, remove his handkerchief, and blow his nose (Merleau-Ponty [1945] 2012: 105–112).

Merleau-Ponty argues that Schneider's behavior reveals a fundamental disturbance in how he experiences space. And this disturbance is the product of Schneider having lost the function of "projection," or of "conjuring up" meaningful situations ([1945] 2012: 115). Maintaining his habitual world of everyday engagements, he nevertheless cannot project other kinds of space, such as the objective space of a coordinate system or the imagined space required for playing or acting. Schneider's loss of the capacity for "projection" reveals two ways that human beings engage with a meaningful world. As Merleau-Ponty argues, in much of our everyday life, we function within the concrete situation of our environment. When I enter the classroom, there is already a network of meaningful relations among myself, my students, the tables and chairs, and the whiteboard. I do not need to actively create this situation. It simply presents itself to me. But we also find occasion to play and act, to speculate, to theorize. And when we perform these activities, we do so within a situation that we ourselves project, or conjure up. Moreover, we do not convince ourselves that this new situation is real. Rather, we move within this situation precisely as fictional, or as imaginary: when I perform Shakespeare's *Hamlet*, I do not take myself to be the real Hamlet; I experience my performance *as* a performance. Likewise, if my friend and I watched The Changing of the Guard in London and I imitate the guard's march, I do not take myself to be the guard; we both experience my behavior precisely as an act of imitation, or play, that only makes sense within the particular imaginative context that we have projected for ourselves. Schneider, and those with similar conditions, cannot have this kind of experience. They

can no longer project new situations around themselves. As Merleau-Ponty says, "The world no longer exists for these patients except as a ready-made or fixed world, whereas the normal person's projects polarize the world, causing a thousand signs to appear there" ([1945] 2012: 115). Schneider cannot even have political or religious opinions—these domains do not reside in his concrete situation and he cannot produce them himself.

Schneider's condition and Merleau-Ponty's analysis have received considerable attention (Dreyfus 2007; Matherne 2014; Romdenh-Romluc 2007). But much of this literature ignores how Merleau-Ponty used this case study as a catalyst and justification for his metaphysical project. Merleau-Ponty suggests that the transcendental idealist—including the Husserlian phenomenologist—could not account for Schneider's condition by appealing to the loss of projection. This is because, for the transcendental idealist, "[o]ne thing alone is comprehensible, namely, the pure essence of consciousness" (Merleau-Ponty [1945] 2012: 127). The transcendental idealist posits an essential, invariant structure of consciousness, and therefore cannot appeal to changes in this structure when trying to make sense of the condition. In light of this metaphysical commitment, he is left with two equally implausible options when presented with the mentally ill person: this person either fails to count as an experiencing subject, or he is deceiving himself—"*Behind* his delusions, obsessions, and lies, the madman *knows that he is* delirious, that he makes himself obsessive, that he lies, and ultimately that he *is not* mad, *he just thinks he is*" (Merleau-Ponty [1945] 2012: 127). A metaphysics of transcendental idealism leaves no room for severe psychopathology because it posits in advance a universal structure of consciousness manifested by all human beings; one either manifests this universal structure or one is not an experiencing subject at all.

In light of this shortcoming, Merleau-Ponty requires a new metaphysics that does justice to the possibility and diversity of mental illness (Fernandez 2015). This project of devising a post-transcendental metaphysics occupied Merleau-Ponty up to his final, unfinished work, *The Visible and the Invisible* ([1964] 1968). However, he set up the problem and provided an initial sketch of his metaphysical solution in *Phenomenology of Perception*. He set himself the task of finding "the means of linking the origin and the essence of the disorder," or of establishing "a *concrete essence* or a *structure* of the illness that expresses both its generality and its particularity" ([1945] 2012: 127). In other words, rather than seek the essence of consciousness as such, he seeks the general structure of each *kind* of consciousness, while also keeping sight of those features that are distinctive of each *particular* case. In so doing, he must acknowledge the

generality of the illness (i.e., those features shared by everyone with this kind of condition) and the particularity of the illness (i.e., those features that are distinctive of the individual's condition). For example, in a study of schizophrenia, the Merleau-Pontyan phenomenologist should articulate general features of schizophrenia (e.g., the general structure of hallucinations), while also attending to its distinctive manifestation in the subject in question (e.g., the particular objects that *this* subject hallucinates).

To develop this new phenomenological metaphysics, Merleau-Ponty aims to establish "a relationship that would be neither the reduction of the form to the content, nor the subsumption of the content under an autonomous form" ([1945] 2012: 128). In other words, we cannot appeal to an autonomous, transcendental structure of experience that determines the content of experience in advance. This wouldn't leave any room for contingent experiences, or the accidents of human life, to alter the fundamental structure of experience—as they do in some cases of psychopathology. According to Merleau-Ponty, we can only resolve this problem if we admit an alternative metaphysical relation between the form and content of experience, one in which "form absorbs content to the point that content ultimately appears as a mere mode of form" ([1945] 2012: 128). He suggests, initially, that this will be a dialectical relationship between form and matter in which the two sides intertwine and mutually shape each other. But this, he admits, is a poor characterization. It assumes that form and content are two independent phenomena that might come into contact. But, in actuality, each is an abstraction from the real phenomenon of "human existence," which he redefines as "the perpetual taking up of fact and chance by a reason that neither exists in advance of this taking up, nor without it" ([1945] 2012: 129).

Merleau-Ponty's approach seems promising, but it's difficult to decipher—in part because he embraces contradictory formulations and paradoxical language. We can clarify his approach by drawing a parallel with Gestalt psychology, which had a major influence on his thinking. The Gestalt psychologists argued that our experience is self-organizing. I do not impose a form on the contents of my experience. Rather, the contents of experience form themselves into a self-organized, meaningful whole. It is always this whole that I am presented with, which is not simply a sum of distinct parts. Merleau-Ponty, inspired by this psychological insight, transformed it into a metaphysical insight about human existence itself. Human existence does not have some necessary form that is simply filled in by contingent content. Rather, what we call the "form" or "structure" of human existence is a description of the holistic organization that my experience and existence assume. Because this form is not independent or

autonomous, it can be fundamentally reshaped by the content of life-experiences. Human existence is always shaped by its natural and cultural circumstances. This is not to say that the subject is passively constituted by her environment. She also constitutes the sense and meaning of her environment in turn. Merleau-Ponty's point is to highlight that there is no autonomous, unchanging form of human existence that underlies our particularities and differences. Insofar as we are subject to the circumstances of life, we are radically contingent beings.[3]

In this formulation of human existence we find the metaphysical possibility of mental illness, or of the disordered subject. According to Merleau-Ponty, it is because our subjectivity is embedded in a material body and a concrete environment that we can become ill in the first place.

Metaphysics: Contemporary Applications

Of course, our goal is not simply to *understand* Merleau-Ponty's new metaphysics—our goal is to apply it. But how does one "apply" a metaphysics? Metaphysics seems the most abstract of philosophical inquiries, concerned with the nature of reality, not the practical study of mental disorders. Yet, the metaphysical system that we subscribe to always comes with a host of assumptions about our subject matter: What we choose to study, how we approach our objects of investigation, and how we interpret our findings are determined by our metaphysical presuppositions. As Merleau-Ponty points out, if a psychologist subscribes to transcendental idealism then he cannot take the possibility of mental disorder seriously. Rather than explain it, he explains it away; it has no place within his metaphysical system.

A phenomenologist who takes up Merleau-Ponty's metaphysics will interpret her subject matter differently; she will open herself to possibilities that might be neglected by a more traditional phenomenological metaphysics because she will not presume the necessity or invariance of any particular structure of subjectivity. We already found an example of this in Merleau-Ponty's study of Schneider, where he argues that Schneider's condition is best understood not as the projection of a distressing situation (as we find in neuroses), but as the loss of the capacity to project such a situation in the first place. This possibility, he suggests, is not open to the transcendental idealist because it conflicts with what one might take to be the pure essence of consciousness. It would have to be dismissed as a metaphysical impossibility. The transcendental idealist, confining himself to those accounts that make sense within his metaphysical framework, is left with a much smaller range of possible alterations to appeal to.

How does this model play out in contemporary phenomenological psychopathology? Many contemporary phenomenologists still subscribe to some form of transcendental idealism, insofar as they take on some of the metaphysical commitments of Husserl or early Heidegger. Even in contemporary studies of psychopathology, phenomenology is often characterized as the study of *essential* structures of consciousness, which seems to draw an in-principle distinction between those structures of consciousness that can change and those that cannot.

We can illustrate how these metaphysical presuppositions play out in two examples: affectivity in major depressive disorder (MDD) and selfhood in schizophrenia. Phenomenologists often rely on first-person reports, whether from qualitative studies, memoirs, or face-to-face conversations. The phenomenologist examines these first-person reports, and then articulates how the structures of experience must have altered such that the world would be experienced in the way the subject describes it.

In the case of MDD, many people report changes in their moods and emotions. As a result, phenomenologists (and psychiatrists) often characterize depression as a distinctive mood, or way of being affectively attuned to the world—e.g., profound sadness, grief, guilt, or despair. Many of these phenomenologists also work within an early Heideggerian framework, which, like the Husserlian framework, assumes a level of essential, invariant structures of human experience. Heidegger, for instance, argues that affective attunement is essential to human existence and that human beings are therefore always attuned to their world through some mood or other; it is impossible to be without a mood (Heidegger [1927] 1962: 172–179). However, many people describe their depression as a loss of moods and emotions *in general*, rather than as a certain kind of mood, or even as a loss of particular moods, such as happiness or joy (Fernandez 2014). In his memoir, *The Noonday Demon*, Andrew Solomon describes this experience:

> The first thing that goes is happiness. You cannot gain pleasure from anything. That's famously the cardinal symptom of major depression. But soon other emotions follow happiness into oblivion: sadness as you know it, the sadness that seemed to have led you here; your sense of humor; your belief in and capacity for love. Your mind is leached until you seem dim-witted even to yourself […] You lose the ability to trust anyone, to be touched, to grieve. Eventually, you are simply absent from yourself. (Solomon 2001: 19)

Reports like this—where one cannot feel sadness, cannot even grieve—pervade the literature on depression. Yet, in both the phenomenological and the psychiatric literature, these reports are often not explained; they are explained

away. The reported *loss* of feeling is redescribed as a particular *kind* of feeling (e.g., a *feeling* of not feeling). At least in the phenomenologist's case, this tendency to redescribe a loss of feeling as a kind of feeling falls in line with Husserl's and Heidegger's metaphysical presuppositions. There are certain structures of experience that we presume to be *essential*. If we know in advance that they cannot be lost, then we need not consider such a possibility.

A similar debate plays out in the phenomenological literature on selfhood in schizophrenia and other disorders. Husserlian phenomenologists are generally committed to the claim that some sense of selfhood, at least in the minimal sense of "for-me-ness," always accompanies experience. This aspect of minimal selfhood is part of the essence of experience; it would be impossible to have an experience without it. But, as Dan Zahavi points out, some philosophers argue that psychopathology provides an objection to this account (Zahavi Forthcoming). They argue that some experiences, such as thought-insertion and severe depersonalization—both symptoms of schizophrenia—are characterized by an experience without a basic sense of for-me-ness, or sense of mineness. Zahavi argues, in contrast, that if we adequately clarify our various notions of selfhood, we will find that even in these psychopathological cases a sense of for-me-ness is retained, necessarily entailing that some sense of self and self-awareness remains intact.

There is, however, an important point to consider in Zahavi's approach. While he might begin his investigations within a Husserlian framework, he remains open to the possibility of deep contingencies and alterations in the structure of experience. He does not simply *assume* that we retain a minimal sense of self even in cases of thought-insertion and severe depersonalization. He carefully considers the evidence and provides arguments for his position. This, in the end, is not so different from Merleau-Ponty's approach to the study of psychopathology. While Merleau-Ponty would likely admit more contingency in the structure of experience than would a Husserlian phenomenologist, he sometimes makes essentialist claims about experience, and seems willing to defend them—e.g., that the perception of a figure is possible only if the figure is presented upon a background.[4]

By providing these examples, I am not arguing that it is necessarily incorrect to characterize the affective dimension of depression as a kind of mood, or to characterize experiences of thought-insertion and severe depersonalization as retaining a basic sense of selfhood. Rather, I want to illustrate how our metaphysical commitments predetermine how we interpret our evidence—what we dismiss, what we take seriously, what we feel justified in redescribing,

and so on. This holds for evidence beyond self-reports, but contemporary phenomenological psychopathologists often rely on self-reports as a major source of evidence, so it is worthwhile to think about how we engage—and how we should engage—with such evidence.

Merleau-Ponty's metaphysical commitments—in contrast with Husserl's and the early Heidegger's—allow more leeway in how we interpret our case studies. And Merleau-Ponty developed this metaphysical stance precisely because he was unsatisfied with classical phenomenology's inability to do justice to the phenomena. For Merleau-Ponty, we always need *some* metaphysical system to ground and guide our investigations. But this metaphysical system is only as good as the results it produces. As soon as it forces us to make absurd claims, to dismiss evidence, or to devise unjustified interpretations of the phenomena, it is the metaphysical system—and not the phenomenon itself—that we must dispense with.

Method

Despite differences among phenomenologists concerning the right metaphysical foundations for understanding human existence, there are broad methodological commitments that all phenomenologists share. One of these is the methodological commitment to critically evaluating and suspending our presuppositions and prejudices. If we do not critically evaluate our presuppositions, then we risk describing our phenomenon of investigation not as it is, but as we already believe it to be. Rather than accurately describing what Husserl calls "the matters themselves," we might simply reiterate the current cultural or scientific conception of our subject matter. But while all phenomenologists agree that we should maintain a critical attitude toward our prejudices, they do not agree on how we should evaluate and suspend these prejudices. They employ different methods of evaluation and suspension, and they express varying degrees of optimism about how successful these methods can be.[5]

Husserl's Reduction

In Husserl's case, the central element of his method is the phenomenological reduction. Depending on how we interpret it, there are either multiple reductions—the transcendental, the eidetic, and so on—or multiple stages of the reduction. Which interpretation we adopt here is not especially important, but

our focus will be on the epoché, which we might characterize as either the first reduction or the first stage of the reduction, depending on the interpretation we employ.

Husserl borrowed the term "epoché" from the Greek skeptics, who argued that we should suspend judgment on any matter until we have considered all of the evidence. But for Husserl, the epoché isn't so much a suspension of explicit judgments as it is a suspension of our tacit beliefs about the world and our experience of it. Moreover, Husserl's epoché is first and foremost a change in attitude—a shift from the natural attitude to the phenomenological attitude. The natural attitude is the attitude of everyday life in which we take our experience of the world for granted; we do not reflect on how the world is presented to us through experience. This everyday attitude is problematic for the phenomenologist because she aims to articulate how we experience the world and how our subjectivity plays a role in its presentation. She therefore shifts herself into a phenomenological attitude, focusing on *how* we experience, rather than *what* we experience. As Robert Sokolowski puts it, in the epoché "[w]e look *at* what we normally look *through*" (Sokolowski 2000: 50).

This shift into the phenomenological attitude requires that we suspend, or bracket, our host of presuppositions and prejudices about the world. For example, to attend to how an object of experience is presented to me—say, my coffee mug—I have to suspend my beliefs that the mug is already there, existing independent of me, taking up physical space, available to others, and so on. As a phenomenologist, I ask *how* the coffee mug presents itself to me *as* existing independent of me, *as* taking up physical space, *as* available to others, and so on. When producing these descriptions I cannot presume, for instance, that the world is already populated by physical objects and that my perceptual capacities merely present me with appearances of these physical objects. Even if this were an accurate natural-scientific account of perception, it is not the kind of account the phenomenologist is after. It fails to describe *how* these objects are presented to me in the first place.

While Husserl remained committed to the importance of the epoché, his method evolved over his philosophical career. When his interest shifted from the necessary and universal structures of experience to the contingent and culturally relative, he developed new methods to accommodate his new subject matter. In his studies of the cultural constitution of experience, Husserl realized that our cultural prejudices shape our experience to such a profound degree that we cannot simply suspend these prejudices in one fell swoop. They are ingrained in us, shaping not just our beliefs about what we experience, but even *how* we

experience. Therefore, when we try to enter the phenomenological attitude via the epoché, we often retain prejudices that we didn't even know we had.

This poses a problem for the phenomenologist, who wants to return to "the matters themselves." If our presuppositions about our subject matter are so ingrained that we cannot bracket them through the epoché—i.e., through a shift in attitude—then how can we continue to do phenomenology? If we are incapable of extricating ourselves from our own prejudices, then we are condemned to describe what we already believe to be the case, and nothing more. In light of this problem, Husserl takes a methodological turn, although his goal remains the same. Rather than bracket his presuppositions through a change in attitude, he brackets them by actively seeking them out and analyzing them. Only through this process can he loosen his presuppositions to such a degree that he might attend to the matters themselves.

Some scholars argue that Husserl's late works usher in a different kind of epoché, what David Carr calls the "historical reduction" (2009). This new epoché aims at the same ends—the loosening and suspending of prejudices—but employs the means of historical inquiry. If we want to understand our current prejudices and how they shape our experience, we need to discover the origins of these prejudices and find where they sedimented into our cultural lifeworld—i.e., the meaningful world of everyday life. To take one example, consider Husserl's historical inquiry into Galileo's "mathematization of nature" (Husserl [1954] 1970). He argues that Galileo's major contribution to scientific thought was not his discovery of basic principles of the physical universe, but his reconceptualization of nature *as objective*. He claims that prior to this reconceptualization, we experienced the natural world as shared, as available to others in the way it is available to us. But we did not experience it as objective, in the sense of some independent reality existing behind the merely subjective appearances presented in our experience. It was Galileo's novel conception of nature, he argues, that sedimented into our lifeworld—at least the European lifeworld—and became the normal way of experiencing our environment. According to Husserl, Galileo initiated a worldview in which we discount our everyday experiences in favor of the insights of a particular cultural practice—the practice of natural science.

While many celebrate this worldview, Husserl laments the fact that this Galilean insight covered over the richness of everyday experience. It is this experience that phenomenologists aim to describe, but that a natural-scientific worldview ignores. To get back to this experience, he argues, we must discover the prejudices that have reshaped our experience, and only then can we suspend

them and return to "the matters themselves." It is this new method, I argue, that inspires Merleau-Ponty's approach to psychopathological experience and subjectivity.

Merleau-Ponty's Incomplete Reduction

Merleau-Ponty famously said, "The most important lesson about the reduction is the impossibility of the complete reduction" ([1945] 2012: lxxvii). This is not a dismissal of Husserl's reduction. Merleau-Ponty does not believe the reduction is impossible, only that its *completion* is impossible. And this falls in line with Husserl's late approach, which requires that we remain open to the possibility of being misled by presuppositions and prejudices that we didn't know we had.[6] This is because we cannot fully extract ourselves from our embodied, historical context:

> This is why Husserl always wonders anew about the possibility of the reduction. If we were absolute spirit, the reduction would not be problematic. But since, on the contrary, we are in and toward the world, and since even our reflections take place in the temporal flow that they are attempting to capture … there is no thought that encompasses all of our thought. Or again, as the unpublished materials say, the philosopher is a perpetual beginner. This means that he accepts nothing as established from what men or scientists believe they know. (Merleau-Ponty [1945] 2012: lxxvii–lxxviii)

We must subject both our everyday and scientific beliefs to critical scrutiny. And, whenever we are able, we must hold these beliefs in abeyance until we consider all of the available evidence and properly attend to our phenomenon of investigation.

But what does this have to do with psychopathology? When Husserl developed his historical epoché, he was concerned with how geometry and post-Galilean physics altered the lifeworld of scientists, as well as the lifeworld of everyday people. He was interested in how scientific advancements prompted us to reconceive the natural world as *objective* and, as a result, doubt our own experiences insofar as we conceive of them as *subjective*. Merleau-Ponty carries on this tradition, but with an eye toward the prejudices of the contemporary cognitive sciences, including psychology, psychiatry, and neuroscience.

We find one of the best examples of this approach in Merleau-Ponty's introduction to *Phenomenology of Perception*, entitled, "Classical Prejudices and the Return to Phenomena." Here, he undermines the prejudices and presuppositions that we might have about our subject matter—in this

case, perception—to get at the "phenomena," or what Husserl would call "the matters themselves." But Merleau-Ponty does not ask us to slip into a phenomenological attitude in which our "classical prejudices" fall away, leaving us with the untainted phenomena. Rather, he undermines prejudice by confronting it head on. He carefully analyzes standard concepts—including "sensation," "association," "memory," "attention," and "judgment"—to show how our typical understanding of these concepts misleads us, causing us to misconstrue the nature of perception. In his study of "sensation," for example, he shows that the three competing theories—sensation as (1) impression, (2) quality, and (3) immediate consequence of a stimulation—actually share a problematic presupposition. They presume that perception is built up of little bits of sense data that the subject conglomerates into a holistic experience (even if they disagree on the nature of this sense data and how the subject conglomerates them). The trouble with these views is that they explain the *act* of perceiving by appealing to *objects* of perception; they neglect the *how* in favor of the *what*. In short, they beg the question—they do not tell us *how* we experience objects *as* objects because they presume the existence of experiential objects (i.e., bits of sense data) in their explanations. As Merleau-Ponty says, "We thought we knew what sensing, seeing, and hearing are, but now these words pose problems. We are led back to the very experiences that these words designate in order to define them anew" ([1945] 2012: 10). What Merleau-Ponty provides us with in his introduction is not a new theory of perception, but a Socratic inquiry that makes manifest just how little we know about perception. Only by engaging in this negative inquiry, and discovering what we do not know, can we put ourselves in a position to develop a new and better theory of perception. This process, of course, can never be completed. There will always be more presuppositions and prejudices that we have missed. And we are therefore left with Merleau-Ponty's incomplete reduction.

Method: Contemporary Applications

The application of method—unlike the application of metaphysics—is fairly straightforward: to apply a Merleau-Pontyan reduction to the domain of psychopathology, we should critically evaluate the basic concepts and presuppositions that ground this field of inquiry. If we fail to do this, we risk obscuring the psychopathological phenomena by describing what we already believe to be the case.

To illustrate this, we can consider examples of depression and bipolar disorder. As Giovanni Stanghellini (2004) points out, recent editions of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*) and the *International Classification of Diseases* (*ICD*) insufficiently describe mental disorders and fail to clearly delineate their boundaries. Consider one of the central diagnostic criteria for MDD: depressed mood. Is a depressed mood a kind of mood or a loss of feeling? Is it qualitatively different from a normal, healthy mood, or is it simply a matter of degree? Is it a pathological form of sadness or is it different from sadness? None of these questions are answered—or even addressed—by contemporary diagnostic classifications. This is not because the answers are exceedingly difficult to come by, but because it is presumed that we all know what we mean by "depressed mood." We do not even realize it is a question that needs asking. But as phenomenologists have shown, when it comes to our own experience we know considerably less than we think.

A similar problem plays out in the psychiatric literature on bipolar disorder. The term "bipolar" suggests that the condition not only includes two distinct states—depression and mania—but that these states are polar opposites. This feature of the condition is contained in its label, and the term therefore predisposes us to think about the condition as a movement between these two poles. But if we look to the history of the concept of bipolar disorder, we find that the bipolar model—which succeeded the concepts of "manic-depressive insanity" and "manic-depressive reaction"—was just one of three competing models. Other researchers put forward the "triangular model" and the "continuum model." On the triangular model, there are three distinct states—depression, mania, and euthymia (i.e., a healthy, positive mood)—that are related like points on a triangle. This means that one can move from one state to any other without passing through the third state. One advantage of this model is that is allows us to explain mixed states—i.e., states in which symptoms of depression and mania manifest simultaneously—as a shift from depression to mania, or from mania to depression, that does not pass through euthymia. On the continuum model, the two poles are actually euthymia and mania, with depression in the middle. In this case, depression and mania are not conceived of as opposites, and mixed states are unsurprising because any move between euthymia and mania must pass through a state of depression. Importantly, there does not seem to be much evidence that supports the bipolar model over the other two models. For one reason or another the bipolar model caught on and is now the established label for this condition. In forgetting the history of this concept, it is all too easy to begin our investigations from the

assumption that depression and mania are, in fact, polar opposites, and this can problematically predetermine how we describe the experiences of depression and mania (Fernandez 2016).

Leaving aside the philosophical inquiry into "the matters themselves," why should psychiatrists and those living with mental disorders be concerned about undefined or poorly defined diagnostic criteria? If psychiatrists already have a basic sense of what they mean by "depressed mood," isn't that enough for clinical practice? One problem is that there is no reason to think that psychiatrists agree on their definitions of diagnostic criteria. If this is the case, then they certainly do not agree on how and when to apply these criteria. Perhaps one psychiatrist believes that depressed mood is like severe sadness, another thinks it is a deep sense of guilt, and yet another thinks it is a loss of feeling. These psychiatrists will hardly be able to agree on who should be diagnosed with MDD, which kinds of interventions should be made for which individuals, and when to deem someone cured or healthy. And this is only in the domain of clinical practice. If we delve into psychiatric research we will find a host of new problems, such as disagreements in selecting who should be included in a study or in how to assess the efficacy of an intervention.

But this risks confusing two important issues. Some of these problems might be overcome by stipulating clear definitions of each diagnostic criterion and, therefore, clear definitions of each category of disorder. This would at least produce agreement among psychiatrists—they will diagnose the same patients in the same way, establishing what psychiatrists call "reliability." However, phenomenologists should not be satisfied with clearly stipulated definitions. They should strive for definitions that get at the heart of the matter, clearly articulating the nature and boundaries of the condition in question. They should want their definitions and descriptions of disorders to be "valid."[7] When it comes to "depressed mood" or "bipolar disorder," the phenomenologist should not want the psychiatric community to simply put forward a definition. Rather, we should carefully attend to the experience that we are trying to point to with each criterion, discover whether it is in fact a unitary phenomenon (or perhaps diverse phenomena clumped under a common label), and articulate its essential features. Only by proceeding in this fashion can we ever hope to fully understand the psychopathological condition in question.

To enact Merleau-Ponty's reduction, or epoché, phenomenologists need to attend to the concepts used in contemporary psychiatric research and practice, as well as in everyday discourse, and unearth what we are trying to refer to with these terms. In many cases, we might discover that those who

employ the term do not have a good sense of its meaning or its referent. Such an inquiry does not reveal the nature of the phenomenon in question, but it at least reveals how little we know, and thereby puts us in a better position to explore and articulate our phenomenon of interest. Only after interrogating our basic concepts will we be in a position to bracket out our presuppositions and proceed in our phenomenological investigations of the psychopathological condition itself.[8]

Conclusion

The application of phenomenology to the study of mental disorders is by no means new, having been established at least as early as Jaspers's *General Psychopathology* (Jaspers [1913] 1997). It is, however, undergoing resurgence today, especially in the philosophical and psychiatric literature on schizophrenia, depression, and various neurological disorders. But much of this work, both historical and contemporary, applies classical phenomenological concepts—such as selfhood, affectivity, and temporality—without critically engaging the phenomenon of mental disorder itself. There is a general assumption that insofar as phenomenology inquires into human experience, it is equipped to investigate *any* human experience. As I have shown, however, the traditional Husserlian approach is not well equipped for the study of dramatic shifts from normal to pathological experience. Husserl's metaphysics presumes an essential structure of experience that is not susceptible to alteration. And his early version of the epoché overestimates our ability to suspend our presuppositions and prejudices (although his later method moves in the right direction). Merleau-Ponty, by contrast, offers a phenomenological approach that takes psychopathology seriously. He devises a new metaphysics that allows for contingency in the most fundamental structures of experience and existence. And his method acknowledges our human limitations, such as our inability to fully extricate ourselves from long-held beliefs. He therefore recommends critical engagement with contemporary scientific concepts; only by discovering their shortcomings can we put them out of play and return to the phenomenon anew. Through his commitment to accommodating the contingencies of human life, Merleau-Ponty—more than any other phenomenologist—provides the metaphysical and methodological foundations for the contemporary study of psychopathology.

Acknowledgments

I would like to thank Robin Muller, Brad Peters, Gina Zavota, and the participants in the May 29, 2017, Science and Technology Studies Working Group at University of King's College, Halifax, for their helpful comments on earlier drafts of this chapter. I would also like to thank the Killam Trusts for funding this project.

Notes

1 This contrast between Sartre and Merleau-Ponty is a matter of emphasis, rather than a strict divide between their interests and concerns. Moreover, it relies on a caricature of Sartre's thought—albeit one that Merleau-Ponty himself perpetuated in some of his work. For a careful account of Sartre and Merleau-Ponty's debate over the nature of freedom—and Beauvoir's mediation—see Wilkerson (2010).

2 There are, of course, some exceptions. Binswanger and Boss, for instance, seriously considered the relation between their phenomenological approaches and biomedical models of psychopathology. Merleau-Ponty even drew on certain elements of Binswanger's metaphysics, but I cannot go into this here.

3 There are a host of problems that arise from Merleau-Ponty's radical contingency, but I cannot address them here. Perhaps the most important is the difficulty of reconciling radical contingency with Husserl's critiques of psychologism and historicism—i.e., the philosophical claim that mathematics and logic are relative to the structure of the human mind, or to particular cultural configurations, and are therefore only true for us. Husserl's transcendental ego was meant to address this problem by establishing essential structures of experience. But Merleau-Ponty's embrace of the radically contingent subject seems to dispense with this solution to psychologism. For an account of how Merleau-Ponty reconciles his account of human subjectivity with the possibility of mathematical truths, see Hass and Hass (2000).

4 Some interpreters grant considerably more weight to Merleau-Ponty's apparently essentialist claims, and have used these claims to argue that Merleau-Ponty is himself a transcendental philosopher. For an example of this position, see Sebastian Gardner (2015).

5 For accounts of the various roles that prejudices play in phenomenological research, see Fernandez (2016, 2017).

6 This general stance seems to be shared by most post-Husserlian phenomenologists, including the hermeneutic phenomenologists, such as Heidegger and Hans-Georg Gadamer, and the existential phenomenologists.

7 The psychiatric concept of "validity" is notoriously difficult to define, and there is not a currently agreed upon definition (see Tabb 2015).

8 In this respect, the above characterization of applied metaphysics might be interpreted as one segment of applied method. That is, insofar as phenomenologists are methodologically committed to critically evaluating and suspending their full range of presuppositions and prejudices, they are necessarily committed to critically evaluating and suspending their *metaphysical* prejudices (e.g., the prejudices of transcendental idealism). Nevertheless, because of the all-encompassing nature of metaphysical presuppositions, I find it helpful to discuss these presuppositions independently.

References

Carr, D. (2009), *Phenomenology and the Problem of History: A Study of Husserl's Transcendental Philosophy*, 2nd edition. Evanston, IL: Northwestern University Press.

Dreyfus, H. L. (2007), Reply to Romdenh-Romluc, pp. 59–69. In *Reading Merleau-Ponty: On Phenomenology of Perception*. T. Baldwin, ed. London: Routledge.

Fernandez, A. V. (2014), Depression as Existential Feeling or De-Situatedness? Distinguishing Structure from Mode in Psychopathology, *Phenomenology and the Cognitive Sciences*, 13:595–612.

Fernandez, A. V. (2015), Contaminating the Transcendental: Toward a Phenomenological Naturalism, *Journal of Speculative Philosophy*, 29:291–301.

Fernandez, A. V. (2016), Language, Prejudice, and the Aims of Hermeneutic Phenomenology: Terminological Reflections on "Mania," *Journal of Psychopathology*, 22:21–29.

Fernandez, A. V. (2017), The Subject Matter of Phenomenological Research: Existentials, Modes, and Prejudices, *Synthese*, 194:3543–3562.

Fulton, A. (1999), *Apostles of Sartre: Existentialism in America, 1945–1963*, Evanston, IL: Northwestern University Press.

Gardner, S. (2015), Merleau-Ponty's Transcendental Theory of Perception, pp. 29–323. In *The Transcendental Turn*. S. Gardner and M. Grist, eds. Oxford: Oxford University Press.

Hass, M. and Hass, L. (2000), Merleau-Ponty and the Origin of Geometry, pp. 177–188. In *Chiasms: Merleau-Ponty's Notion of Flesh*. F. Evans and L. Lawlor, eds. Albany: State University of New York Press.

Heidegger, M. ([1927] 1962), *Being and Time*. J. Macquarrie and E. Robinson, trans. New York: Harper Perennial Modern Classics.

Heinämaa, S. (2014), The Animal and the Infant: From Embodiment and Empathy to Generativity, pp. 129–146. In *Phenomenology and the Transcendental*. S. Heinämaa, M. Hartimo, and T. Miettinen, eds. New York: Routledge.

Husserl, E. ([1954] 1970), *The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy*, D. Carr, trans. Evanston, IL: Northwestern University Press.

Husserl E. ([1935] 2008), Edmund Husserl's Letter to Lucien Lévy-Bruhl, *The New Yearbook for Phenomenology and Phenomenological Philosophy*, L. Steinacher and D. Moran, trans., 8:349–354.

Husserl, E. ([1913] 2014), *Ideas for a Pure Phenomenology and Phenomenological Philosophy: First Book: General Introduction to Pure Phenomenology*, D. O. Dahlstrom, trans. Indianapolis, IN: Hackett Publishing Company, Inc.

Jaspers, K. ([1913] 1997), *General Psychopathology*, J. Hoenig and M.W. Hamilton, trans. Baltimore, MD: Johns Hopkins University Press.

Kant, I. ([1781] 1999), *Critique of Pure Reason*, P. Guyer and A.W. Wood, trans. Cambridge: Cambridge University Press.

Matherne, S. (2014), The Kantian Roots of Merleau-Ponty's Account of Pathology, *British Journal for the History of Philosophy*, 22:124–149.

Merleau-Ponty, M. ([1964] 1968), *The Visible and the Invisible: Followed by Working Notes*, Evanston, IL: Northwestern University Press.

Merleau-Ponty, M. ([1945] 2012), *Phenomenology of Perception*. Donald Landes, trans. New York: Routledge.

Romdenh-Romluc, K. (2007), Merleau-Ponty and the Power to Reckon with the Possible, pp. 44–58. In *Reading Merleau-Ponty: On the Phenomenology of Perception*. T. Baldwin, ed. London: Routledge.

Sartre, J. P. ([1945] 2007), *Existentialism Is a Humanism*, New Haven, CT: Yale University Press.

Simons, M. A. (1992), Two Interviews with Simone de Beauvoir, pp. 25–41. In *Revaluing French Feminism: Critical Essays on Difference, Agency, and Culture*. N. Fraser and S. L. Bartky, eds. Bloomington: Indiana University Press.

Sokolowski, R. (2000), *Introduction to Phenomenology*, New York: Cambridge University Press.

Solomon, A. (2001), *The Noonday Demon*, London: Vintage.

Stanghellini, G. (2004), *Disembodied Spirits and Deanimated Bodies: The Psychopathology of Common Sense*, Oxford: Oxford University Press.

Tabb, K. (2015), Psychiatric Progress and the Assumption of Diagnostic Discrimination, *Philosophy of Science*, 82:1047–1058.

Wilkerson, W. (2010), Time and Ambiguity: Reassessing Merleau-Ponty on Sartrean Freedom, *Journal of the History of Philosophy*, 48:207–234.

Zahavi, D. (Forthcoming), Consciousness and (Minimal) Selfhood: Getting Clear on for-Me-Ness and Mineness. In *Oxford Handbook of the Philosophy of Consciousness*. U. Kriegel, ed.

8

Transformation through Dialogue: Gadamer and the Phenomenology of Impaired Intersubjectivity in Depression

Constantin-Alexander Mehmel

Introduction

This chapter aims to sketch a phenomenological account of impaired intersubjectivity in depression. By focusing on the relationship between intersubjectivity and depression, specifically on how depression affects but also is affected by interpersonal relations, I intend to shed light on how alterations to our ordinary relations with other people and a presupposed shared background can affect and impair the way we experience the other. The question of if, and to what extent, we can experience the other and have epistemic access to their mind and mental states plays a central role in the current debate regarding the nature and structure of intersubjectivity. It has given rise to what we can broadly identify as the two predominant approaches to intersubjectivity, which in turn provide diverging ways to understand and explain alterations to interpersonal relations in such psychopathologies as depression, autism, or schizophrenia.[1]

On the one side, we have "mentalistic" approaches to intersubjectivity that attribute mental states to the other person to understand their experience. Against the backdrop of what is commonly called the "Theory of Mind" debate, two principal mentalistic approaches have emerged, often referred to as simulationist and theory-theory approaches.[2] Central to both is the assumption that the other's mind is somehow unreachable, which is why they emphasize the need to infer the other's mental states by observing their external behavior. In light of this, pathologies of intersubjectivity are often traced back to impairments in mind reading (e.g., Bosco et al. 2009; Sprong et al. 2007; for an overview, see Hutto 2013).

On the other side, we have "phenomenological" approaches to intersubjectivity that emphasize the affective, embodied, and situated nature of interpersonal understanding (e.g., Fuchs 2013; Gallagher 2012; Ratcliffe 2014a). Although they do not necessarily rule out the contributions of mentalistic approaches in certain cases of interpersonal understanding, their starting point is of an entirely different nature, focusing on the interaction process between two embodied agents. In contrast to an individualistic understanding of the mind, they conceive of the other, including their mental states, as nothing closed-off but, ordinarily, as always already pre-reflectively connected to other people and our environment. This in turn is said to enable an immediate, quasi-perceptual second person experience of the other and their experience, being intentionally directed toward their experience as something distinct and theirs, as the current debates on empathy in relation to intersubjectivity show (e.g., Ratcliffe 2014a; Zahavi 2011). It is on this basis that impaired forms of intersubjectivity in psychopathologies are construed as disturbances to our most basic ways of being with others (e.g., De Jaegher 2013; Fuchs 2013).

The concern of this chapter will not be to assess the ongoing debate between these two paradigmatic approaches to intersubjectivity, however. Rather, I want to turn to a tradition in philosophy and one philosopher's work in particular that has been hitherto neglected in the debate on the structure and nature of intersubjectivity and, as a result, also in the context of impaired forms of intersubjectivity in psychopathologies. In this chapter I intend to show that the insights and theoretical frameworks found in the work of the hermeneutician Hans-Georg Gadamer may be used to shed light on alterations to interpersonal relations in depression. It is interesting that one of Gadamer's principal concepts, which will also prove instrumental to the success of this chapter, namely, "fusion of horizon(s)" (*Horizontverschmelzung*) is regularly referred to for the purposes of advancing diagnostic and therapeutic approaches to psychopathologies (e.g., Fuchs 2010; Stanghellini 2013; Stern 1991; Varga 2015). However, Gadamer's work and hermeneutics more generally has received hardly any attention when it comes to explaining intersubjectivity and pathologies thereof. One notable exception is Amanda Taylor-Aiken, who provides a Gadamerian approach to relatedness and alienation in interpersonal understanding, considering schizophrenia and depression (Taylor-Aiken 2011). While her analysis of depression focuses on alterations to the body, I will offer another potentially fruitful Gadamerian account: by drawing on the fusion of horizon(s), I aim to provide a more general framework for understanding alterations to interpersonal relations to depression that is then also applicable

to alterations to the body. More specifically, it will help us to frame depression as a "dialogical" illness, insofar as it fundamentally alters how we ordinarily relate to other people and the presupposed shared background. I propose that such a framework, highlighting the co-constitution of each other and the space in which we find ourselves, is especially well suited for understanding and articulating different ways of relating to each other—particularly interpersonal relations in depression which I believe to raise both theoretical and practical concerns in the context of pathologies of intersubjectivity in a very illuminating manner.[3]

In what follows, I argue that depression comes with what I call an altered "experience of the other." To understand exactly how depression alters the phenomenology of intersubjectivity, I draw on Gadamer's phenomenology of understanding via the "fusion of horizon(s)" (*Horizontverschmelzung*) as advanced in *Wahrheit und Methode*[4] and thus on his emphasis on transformation through dialogue. I begin by sketching a Gadamerian perspective of an intact dialogue between two people. The rest of the chapter is then dedicated to understanding the deviating forms of dialogue that occur in depression. More specifically, I want to suggest that a Gadamerian approach helps to understand depression as a mode of radical Othering. This allows us to shed light on the specific, yet fundamental impairment in depression—namely, the at least partial, breakdown of the shared, intersubjective horizon in which we ordinarily find ourselves in. This leaves the structure of experiencing the other compromised.

Gadamer on dialogue

To portray an intact experience of the other, thus the phenomenology of intersubjectivity more generally, we need to set out how understanding, involving both cognitive and non-cognitive elements, ordinarily takes place between two people.[5] I suggest that our experience of the other (here synonymous to "another person") is inextricable from coming to understand the other and their claim regarding the mutual subject matter at hand. In other words, a failure in understanding can explain our diminished experience of the other, something key to depression as I will show in the second half of this chapter.

To establish such an intact dialogue, we can turn to Hans-Georg Gadamer and his phenomenology of understanding via the fusion of horizon(s) (*Horizontverschmelzung*). Such an approach has some limitations. Gadamer is

adamant in pointing out that his primary task is of a descriptive nature, setting out how understanding takes place (e.g., *GW2*: 438–439). And yet, precisely by inquiring into the conditions of the possibility of understanding, it soon becomes clear that Gadamer's own phenomenology is normatively laden insofar as it explicates ontological but also epistemically favorable conditions that must be met for understanding to succeed (e.g., *GW1*: 307, 367; *GW2*: 438). In other words, the account of dialogue sketched in this section of the chapter leans on the normative dimension inherent in Gadamer's own work. My proposed Gadamerian perspective of what I call an "intact dialogue" thus amounts to an idealized account of dialogical interaction and intersubjectivity more generally. My analysis is not only normative insofar it suggests how we should interact with one another in the process of coming to understanding each other, but also restrictive in that I am describing only certain domains and ways of social interaction. And yet, I want to suggest that the notion of intact dialogue I advance helps us, first, to bring into focus the deviating forms of dialogue that occur in depression in a contrasting way. Secondly, it allows us to tease out therapeutic implications by pointing toward an ideal dialogical interaction and thus how a therapeutic setting could and, arguably, should look like.

According to Gadamer, the starting point for any dialogue between two people is that each interlocutor enters the dialogue from within a unique horizon.[6] Denoting "the range of vision that includes everything that can be seen from a particular point" (*GW1*: 307; translation modified) a horizon structures one's experience of the other. Being a finite historical being and thus always already situated within a tradition, I bring along certain "tacit expectations of meaning and truth" (Garrett 1978: 393), through which I perceive the other and their claim regarding the subject matter. It would be wrong, however, to understand a horizon as a necessarily restrictive force. Although it does limit our perception of possibilities, it simultaneously provides the conditions whereby we can experience the other in the first place (see *GW2*: 224). So we do not encounter the world at random, but it is already constrained, so to speak, by our prior understanding of what is possible for us, while at the same time this recognizes that such understanding is always open to changes and deviations. A horizon is not closed off, but rather open toward new experiences. As Gadamer puts it, "[a] horizon is not a rigid boundary but something that moves with one and invites on to advance further" (*GW1*: 250). Whenever I experience something new, my horizon is expanding. This openness is indicative of a horizon's essential dialectic, existing only by being superseded (*GW1*: 348). Cast in terms of the self and the other, this suggests that, for Gadamer, both horizons, that of myself

and of the other, are inextricably related. Underlying such openness we can therefore identify the more far-reaching claim that my horizon does not exist independently from the other's horizon, but rather that both belong to a more fundamental, shared horizon (*GW1*: 309).

Gadamer thus appears to advance the phenomenologist Martin Heidegger's notion that we are always already in relation with others, something that is crucial for our project of a phenomenology of intersubjectivity. Although both interlocutors have a unique horizon and thus experience the subject matter differently, they are nonetheless attuned to each other. This holds true regardless of whether or not the different perspectives lead to a disagreement regarding the subject matter. Two people might experience things differently—and in that sense "disagree"—and yet, such divergence is only possible against the backdrop of a presupposed shared background (Ratcliffe 2014a: 272–273). Any dialogue therefore occurs within what we might call a shared, intersubjective horizon in the sense that both parties are already united by it: "I may say 'Thou', and I may refer to myself as over against a Thou, but a common understanding always precedes these situations" (*GW2*: 223). Hence, Gadamer concludes, that the "formulation 'I and Thou' already betrays an enormous alienation," since "there is neither the I nor the Thou as isolated, substantial realities" (*GW2*: 223). Put phenomenologically, we normally take it for granted that we do not experience a situation individually insofar as it would belong to either of us exclusively. Rather, we always already experience it as ours and shared, against the backdrop of which we can then have individual experiences.

We can therefore extract from Gadamer's work the view that any two people conversing with each other do not exist as two isolated realities. Rather, they share in a mutually constituted interpersonal reality, which again is constitutive of their respective outlook onto the world. This wider interpersonal horizon can be understood as a "quasi-transcendental condition"[7] in that, without it, the acquisition of knowledge about the other would be made very difficult if not impossible, and would thus leave the structure of experiencing the other compromised. In other words, such a presupposed shared background makes it possible for the two people entering a dialogue to come to an understanding. Both of their horizons can fuse to what we might tentatively call a "third," more encompassing one, the process of which Gadamer calls "fusion of horizon(s)" (*Horizontverschmelzung*).[8]

However, simply being attuned to each other is not sufficient for what we might call a "successful" fusion of horizons, appreciating the other and their experience as unique and thus theirs. Gadamer emphasizes a fundamental

openness that needs to be present in a dialogue, without which "there is no genuine human bond" (*GW1*: 367). Such mutual openness involves a willingness to be transformed by the other and thus what Gadamer calls the "fore-conception of completeness" (*Vorgriff der Vollkommenheit*), that both interlocutors suppose each other's claim to be meaningful and true (*GW1*: 229). For, only if we deem the other as a possible dialogue partner do we give them enough space to articulate themselves, hence acknowledging them as a person with a unique horizon. Otherwise, we risk projecting ourselves onto the other, whereby we would reduce them to an object-like status and consequently dispense with themselves as a "moral phenomenon" (*GW1*: 364).

From a Gadamerian perspective, we can conclude that an intact dialogue aims at a fusion of horizon(s) with the other, allowing us to experience and thus recognize the other as a person. Intersubjectivity and the experience of the other is not something artificially constructed, "insisting that the Other can first be given only as a perceived thing, and not as a living, as given 'in the flesh'" (*GW10*: 95.). The experience of the other cannot be an act of self-relatedness (*GW1*: 365.), emulating what it is like to be the other from the self's viewpoint. For, this would assume a privileged access to the other's mind (*GW1*: 365.), whereby, however, the experience of the other would be diminished and reduced to a projection of the self. Instead of being open toward the other and immediately recognizing their experience as something distinct and theirs, such an encounter of the other would supersede both the distinction between "my" and "your" experience, and thus between the self and the other.[9]

Key to the phenomenology of intersubjectivity, then, is the mutual recognition of each other as the bearers of unique experiences that can transform us, without which the fusion of horizon(s) will not succeed. In other words, a phenomenology of intersubjectivity, as we have construed it here, involves both the recognition of another person and the resultant fusion of horizon(s). This fusion changes the way both interlocutors relate to each other, as their knowledge of the subject matter but also of the other's view on it has been enlarged. The way one experiences the other has been altered, as one's horizon has been expanded, enabling an experience of the other that was impossible prior to the fusion.

However, this fusion should not just be understood in terms of two individual horizons expanding. For, the prime focus is not on each of the interlocutors and their newly extended horizons, but on the event of the fusion itself. Being mutually open toward each other, they are united by their common aim of understanding the subject matter and thus experiencing the respective other. This event structure can be linked to what Gadamer captures elsewhere with his

concept of "play," a concept often all too neglected in the Gadamer scholarship in this context:

> The primacy of the game over the players engaged in it is experienced by the players themselves in a special way, where it is a question of human subjectivity that adopts an attitude of play … the game itself is a *risk* for the player: one can only play with serious possibilities. … The attraction of the game, which it exercises on the player, lies in this risk. (GW1: 111–112; italics added)[10]

Applied to the fusion of horizon(s), both dialogue partners are guided by the dialogue itself, yielding to an intersubjective dynamic. This is why the fused, "third" horizon constitutes a shared, intersubjective horizon belonging to both rather than either of them exclusively. However, without the willingness to be challenged, thus putting ourselves "into *play* … through being at *risk*" (*GW1*: 304; italics added), we cannot fuse horizons and experience the other. Sketching a Gadamerian perspective of an intact dialogue, we can thus infer that it entails both mutual openness and trust toward the other, without which we cannot appreciate the other and their experience as theirs.

Dialogue in depression

Drawing on Gadamer's phenomenology of understanding via the fusion of horizon(s) has allowed us to sketch a phenomenology of intersubjectivity. We have established how understanding takes place in the context of an ordinary dyadic relationship, that is, between two people, and have thus provided an account of an intact dialogue more generally. Even though our Gadamerian account is of a very preliminary nature, I shall now apply these findings to the phenomenology of depression, elucidating the deviating forms of dialogue that occur in depression and hinting at the transformative potential inherent in the Gadamerian model of dialogue.[11]

It is certainly the case that "depression" is used as an umbrella term for a number of diagnoses with a variety of symptoms. Henceforth, however, I shall employ the term to refer to a phenomenological change in the experience of the other as often to be found in a major depressive disorder, as categorized by the *Diagnostic and Statistical Manual of Mental Disorders: DSM*-5 (APA 2013). In fact, this phenomenological change can be identified in many autobiographical accounts and testimonies of psychiatric clients, all describing an impaired form of intersubjectivity. For instance, consider the following statements:

> Life is about connection. There is nothing else. Depression is the opposite; it is an illness defined by alienation. (Brampton 2008)
>
> When I am depressed I feel like my relationships are less stable and I trust others a lot less. I try to avoid people, as they seem angry and irritated at me. … I feel like a burden. (Ratcliffe 2014a: 274)
>
> I find other people irritating when depressed, especially those that have never suffered with depression, and find the "advice" often given by these is unempathetic and ridiculous. (Ratcliffe 2014a: 274)
>
> I believe depression is—a disease of isolation that tells you to withdraw, stay away, don't be a social person. Stay away from the people who are going to make you better. (Karp 1996: 35)

In these accounts, which I take to be representative of the aforementioned phenomenological change, we can identify the two principal themes of isolation and lack of trust. Interpersonal relations seem, at least most of the time, bereft of any positive, warm dimension. Instead, the depressed experiences the other as a threat and alienating force, with whom they cannot enter a genuine bond. One way of construing this change in experiencing the other, I propose, is in terms of the fusion of horizon(s) between two people, and thus how understanding in an intact dialogue occurs. Whereas a mutual openness lies at the heart of an intact dialogue, a depressed person is lacking such openness in virtue of not trusting the other. As a result, they seem incapable of putting themselves "into play" and "at risk."[12] Not yielding to the intersubjective dynamic of completely letting go in the process of the dialogue, the depressed person prevents themselves from fusing horizon(s) with the other, thus from appreciating the other as a person. Instead, the other is reduced to a projection of the depressed, constituting a threat (see also Ratcliffe 2014b; Styron 1990). We can thus distinguish between two different phenomena: the depressed person is not only reducing the other to a projection of themselves, but also and inextricably relatedly, they are incapable of really interacting with the other.

The lack of trust furthermore explains why other people's advice is deemed "unempathetic and ridiculous." Key to the experience of the other in an intact dialogue is the "fore-conception of completeness," as I have outlined in the first section of this chapter, namely, that both interlocutors suppose each other's claim to be meaningful and true. The depressed person, however, does not seem to be in a position to suppose the other's claim to be meaningful and true, since they have reduced the other to an object-like status of embodying (almost) nothing but threat. The possibility of interacting with the other in a way that could change and transform the depressed person's horizon is diminished. Hence, they do not feel

understood by the other, which in turn makes them feel even more isolated and like a "burden."[13] Even if the depressed person wanted to be understood, "[yearning] for connection," a fusion of horizon(s) could not take place, as they "[are] rendered incapable of being with others in a conformable way" (Karp 1996: 14).[14]

It is plausible to infer that depression involves a diminished experience of the other, more generally an impaired form of intersubjectivity. The account sketched so far reveals the inability to connect and thus experience the other in a horizon-changing way. Without being in a dialogue with the other, however, the depressed person lacks the possibility "of immersion in a dynamic world that incorporates the potential for meaningful change" (Ratcliffe 2014a: 277). Instead, we find the depressed completely shut off from the world:

> On our side of that severed connection, it was hell, a life lived behind glass. (Thompson 1996: 200)

> I feel like I am watching the world around me and have no way of participating. (Ratcliffe 2014a: 277)

An intact dialogue always occurs within a shared, intersubjective horizon that unites both interlocutors. This is why we concluded in the first section of this chapter that the formulation of "I and Thou" does not do justice to our phenomenology of intersubjectivity, as both do not constitute two completely separate realities. The above quotes, however, seem to depart from such an account. Rather than being mutually attuned to each other, I suggest, the depressed person appears to fall out of such a mutual framework. What has been viewed as a quasi-transcendental condition in an intact dialogue; that is, the interpersonal horizon that makes possible the unfolding of the event of interpersonal understanding is missing, which again leaves the structure of experiencing the other compromised, to the effect of affecting how the depression individual experiences themselves in relation to the other.

What this change involves can best be understood, I propose, when broadly conceptualizing the depressed person as what I shall call a "radical Other."[15] As sketched in the first section, in an intact dialogue two people experience things differently in virtue of each having a unique horizon, and yet both belong to a shared, intersubjective horizon. In a dialogue between the depressed and non-depressed person, however, the two perspectives at work differ more fundamentally. For the former does not seem to be part of the same framework as the latter, as we have established before. This is why the depressed person does not feel understood but isolated, feeling completely detached from everyone else without any possibility of taking part in the world, even if physically close. It

thus seems that the impossibility of entering a dialogue cannot be reduced to a purely conscious act on behalf of the depressed person. It certainly is the case that in virtue of the feeling of isolation that the depressed person feels conscious of the impossibility of future engagement, which in turn affects the range of their horizon and thus their future decisions of abstaining from any interaction. Yet, at the same time, such a reading runs the risk of reducing the concept of horizon to purely epistemic terms. A horizon should not be understood as necessarily conscious to either or both interlocutors, but rather in terms of framing the individual's possibilities of engagement with the respective other. It thus becomes clear that because of the, at least partial, breakdown of the shared, intersubjective horizon the depressed person feels the impossibility of being understood.[16]

The lack of a mutually shared backdrop does equally affect the non-depressed person in that they struggle to relate to the depressed person:

> When I start to get depressed, I only filter through the negative messages from friends and family … As a result, they soon learn to step on egg shells around me, they become less affectionate because I'm less receptive. … It's a very hard thing to do to be able to step back and realize that someone who is depressed is projecting their own thoughts onto others. (Ratcliffe 2014a: 279)

The seeming impossibility for the depressed person to fuse horizon(s) thus affects the non-depressed person. Being exposed to sheer negativity, the depressed person is likely reduced to an object-like status, being "unreachable."[17] Such a reduction, however, appears problematic in that the depressed person becomes even more out of reach, if actually being avoided. As Karp states, "the need to withdraw from others overrides the realisation that self-isolation will only deepen one's anguish" (Karp 1996: 35). In other words, through such a reduction and the resultant alienation, we run the risk of dispensing with the depressed person as a moral phenomenon, as another person with unique experiences. This risk is revealing with respect to the phenomenological account of impaired intersubjectivity in depression. Central to the experience of the other is "an appreciation of his potential to reshape one's world" (Ratcliffe 2014b: 236), the potentiality of which the depressed person seems to be lacking in virtue of being "unreachable." Even though the fusion of horizon(s) thus cannot take place, we should nonetheless attempt to "realize that someone who is depressed is projecting their own thoughts onto others" and avoid reducing the depressed completely. For, "[much] of depression's pain arises out of the recognition that what makes me feel better—human connection—seems impossible in the midst of a paralyzing episode of depression" (Karp 1996: 16). Hence, instead of dispensing with the depressed person as a moral phenomenon, our

phenomenological analysis points to the paradoxical situation of the depressed person feeling like a radical Other themselves, and yet ultimately not wanting to be reduced to such. For, on the one hand, it is certainly the case that the depressed person has a diminished sense of agency, feeling isolated and lacking any (or most) interpersonal possibilities, which again gives rise to an impaired form of intersubjectivity. On the other hand, however, it does not take much from here to yield to a reductionist experience of the depressed person, perceiving them as nothing more than an object. This again could amount to a loss of the possibility of helping the depressed person, who is however dependent on our willingness to engage with them in a transformative manner. Unlike what the at least partial breakdown of a shared, intersubjective horizon may suggest, the open and dynamic nature of everyone's unique horizon entails, or at least I propose that it entails, the very possibility of re-entering this mutual framework. Instead of reducing the depressed person to an object-like status, the non-depressed person should continue to practice openness and trust.

Although I cannot conclusively argue for this thesis in this chapter, I believe that any form of therapeutic encounter between the depressed and non-depressed person would immensely profit from a Gadamerian approach to dialogue and intersubjectivity more generally. A willingness to engage with the depressed person and to allow oneself to be put at risk by the depressed person, thus being potentially challenged in one's own assumptions (*Vorurteile*), would likely generate an environment in which mutual openness and trust once again become possible. A Gadamerian approach suggests that understanding the other and thus empathizing with them is a mutual, interactive and co-constituted process, a to-and-fro movement between two people. The other empathized-with is nothing static but someone "that moves with one and invites one to advance further" (*GW1*: 250), insofar as the self's experience of them is continuously changing and being changed in the never-ending, never-final fusion of horizon(s). A long-term willingness to engage with the depressed person would thus likely help to transform the depressed person's experience of themselves and of the other, more generally generate an environment in which transformation could occur.[18]

Conclusion

Our aim in this chapter has been to sketch a phenomenological account of impaired intersubjectivity in depression. Our claim has been that drawing on Gadamer's phenomenology of understanding via the fusion of horizon(s) helps elucidate how depression affects the phenomenology of intersubjectivity.

Against the backdrop of an intact dialogue between two people, we have construed the deviating forms of dialogue that occur in depression in terms of the seeming impossibility of fusing horizon(s). Even if the account given here does not fully apply to all cases of depression and furthermore might not be applicable to other kinds of psychiatric illnesses, I hope that the reader will nonetheless realize that such a hermeneutically inspired phenomenological sketch enables an understanding of depression as a mode of radical Othering that might otherwise not be possible. Overall, I hope to have advanced a way to conceptualize experiences of depression that is of help for philosophers and non-philosophers alike.

Acknowledgments

Earlier versions of this chapter were presented at Durham University, Cardiff University, the University of Warwick (Coventry), Central European University (Budapest), and Vita-Salute San Raffaele University (Milan). I am grateful to participants of all events for their comments, in particular Anna Bortolan, Ian James Kidd, and Jonathan Webber. I also wish to thank Harry Lewendon-Evans and Johannes Roessler for their insightful remarks and Louise Campion, Jack Kellam, and Thomas Stephen for their careful reading and editing. Finally, I want to thank Matthew Ratcliffe for introducing me to the phenomenological approach to the understanding and treatment of psychopathologies.

Notes

1 The subsequent overview follows Fuchs (2015).

2 For simulationist approaches, see, for example, Gordon (1986), Heal (1996), and Goldman (2006). For theory-theory approaches, see Carruthers and Smith (1996) and Stich and Nichols (2003).

3 In advancing this Gadamerian hermeneutic approach to impaired intersubjectivity I also hope to begin in addressing the discrepancy between the role of hermeneutics in sketching an account of intersubjectivity, on the one hand, particularly in the context of psychopathologies of intersubjectivity, and diagnostic and therapeutic approaches to those psychopathologies, on the other hand.

4 All in-text page citations in this chapter refer to *Gesammelte Werke* vol. 1 (2010), though I am quoting (unless otherwise noted) from the translation Gadamer (2004). Hereafter referenced as GW1. All the other in-text references to Gadamer's

body of work will be given in the same fashion, referring to *Gesammelte Werke* vol. 2 (1993), hereafter referenced as *GW2*, and vol. 10 (1995), hereafter referenced as *GW10*, though I am quoting (unless otherwise noted) from the following translations cited in the reference list: Gadamer 1976, 1989, 2000, and 2006.

5 The Gadamerian notion of understanding is a rather extended one that is more than the cognitive "faculty" and inclusive of non-cognitive elements such as certain non-conscious prejudices making up our pre-understanding of the world. This is why, as it will become clear throughout the chapter, that, for Gadamer, the event of understanding cannot solely be construed in epistemological terms, coming to know something and making an intellectual judgment, but rather as affecting how we find ourselves in the world more generally. As he puts it so poignantly, "understanding in a dialogue is not merely a matter of putting oneself forward and successfully asserting one's own point of view, but being transformed into a communion in which we do not remain what we were" (*GW1*: 384; italics added).

6 Shortly, this uniqueness is qualified by being understood against the backdrop of an underlying shared horizon. It should not be taken to indicate either a degree of isolation or relatedly the possibility of understanding in abstraction from interaction with others.

7 While Gadamer aligns himself with Kant's transcendental project of exposing *a priori* structures of understanding (see Veith 2015: 12), here it makes more sense to frame it as a quasi-transcendental condition. For, if it is construed as a transcendental condition, it is the condition for the possibility of encountering whatever the other is at all, before we even get to propositional knowledge. However, it is certainly the case that I might be able to have an outer perception of the other's bodily expressive behavior and thus understand that they are experiencing something, without being part of the same wider, intersubjective horizon. The latter only seems to be a requirement to gain more than a vague awareness of the other's mental states. For a nuanced approach to intersubjectivity that accounts for different levels of empathic relating to the other, see Stein's three-tiered account of empathy (Stein 2008).

8 The German "Horizontverschmelzung" is a somewhat ambiguous singular, which I decided to render as "fusion of horizon(s)" in English, thus alluding to the plural singularity as key to the process: fusing unique horizons against the backdrop of an underlying, unifying horizon.

9 The more general difference between approaches open toward the other and emulating the other can also be cast in non-Gadamerian terms as one between phenomenological and simulationist approaches to empathy, broadly echoing the distinction between phenomenological approaches and mentalistic approaches to intersubjectivity. For an overview and analysis of the extent to which those overlap, see Ratcliffe (2012, 2014a), or Zahavi (2011). As will become clear, however, particularly in our context, the here presented Gadamerian approach toward the

other, and to intersubjectivity more generally, diverges from the phenomenological approach to empathy in at least one crucial respect: it does not suffice to be intentionally directed toward the other and their experience. For, to truly recognize the other as a bearer of unique experiences, we need to be aware of the possibility that we might not be sharing a mutual, intersubjective horizon, which is to say that their experience might be rather different from ours and beyond our own everyday way of being in the world. In a sense, this chapter points toward what we might call a hermeneutic approach to empathy in relation to intersubjectivity, highlighting the co-constitution of oneself and the space we find ourselves in.

10 In Gadamer scholarship, "play" is often reduced to its role for our experience of art, whereas Gadamer's notion of "fusion of horizon(s)" is used to make sense of Gadamer's model of understanding more generally. Ignoring the subtleties of Gadamer's position, such a reading preserves the focus on the unfortunate metaphor of fusing horizon(s), which has been the prime target of critique see Mehmel (2016). For an exception, see, for example, Vilhauer (2009, 2010). To avoid such a danger, this chapter employs the concept of "play" in relation to and thus as a corrective of the concept of fusion of horizon(s).

11 In this context, I understand dialogue as broad as encompassing any communicative interaction between two people. In contrast to Gadamer who ultimately conceives of dialogue in purely linguistic terms and neglects the bodily dimension more generally (for exceptions, see, for example, *GW2*: 208, 335–336), such a definition can account for the fact that in a dialogue two embodied agents interact. While it is beyond the scope of the chapter, a comprehensive discussion of impaired intersubjectivity in depression would have to acknowledge the alterations to how the depressed person experiences their body as well as that of the other (see, for example, Fuchs 2013). In lieu with the chapter's operating metaphor of the fusion of horizon(s), we could frame it as the seeming impossibility of a bodily dialogue, thus of fusing the bodily horizons of the depressed and non-depressed person (see Taylor-Aiken 2011). In fact, the fusion of horizon(s) could be understood in different terms—e.g., cognitive, bodily, affective—the distinction of which would allow for a more sophisticated understanding of depression and its various manifestations.

12 This seeming incapability of putting oneself "into play" could be further substantiated by drawing on Gadamer's account of language as advanced in the third and final part of *Wahrheit und Method*. This would include recognizing the inability to communicate one's depression as a core and painful aspect of the depressed person's experience. For Gadamer, understanding always already takes place in language, which constitutes the foundational horizon for our hermeneutic experience as such: "Our first point is that the language in which something comes to speak is not a possession at the disposal of one or the other of the interlocutors. Every conversation presupposes a common language, or better, creates a common

language" (*GW1*: 384). In the dialogue between the depressed and non-depressed person, however, this common language can no longer be necessarily presupposed. Struggling to engage in seamless linguistic play, we can infer, the depressed person is prevented from creating a common language with their interlocutor and thus becomes even further isolated. Whereas in an intact dialogue language can function as "a medium where I and the world meet, or, rather, manifest their original belonging together" (*GW1*: 478), the depressed person feels detached from the world. Gadamer's conception of the fundamental status of language for our experience of and in the world thus allows emphasizing something like a linguistic separation between the two interlocutors as opposed to the linguistic unity at work in an intact dialogue.

13 Whether or not the feeling of isolation precedes the feeling of not being understood, in my view, does not have any bearing on the here presented Gadamerian reading. In fact, I would rather want to point to the relation between both phenomena, insofar as it seems that the feeling of not being understood cannot be untangled from the depressed person's own negative feelings.

14 This raises an interesting question: To what extent does the feeling of not being understood correspond to the other person's inability to understand? For, insofar as the feeling of not being understood and the feeling of isolation are inextricably related (see previous footnote), the former might possibly be the result of the latter or at least exacerbated thereby, regardless of whether or not an at least partial understanding might have been achieved.

15 This notion and its implications are inspired by Emmanuel Lévinas's radical alterity (e.g., Lévinas 1969: 194) and Sara Ahmed's "affect aliens" (Ahmed 2010).

16 Here, I am not so much concerned with introducing a temporal order of what comes first. Rather, it is important to distinguish between an ontological and epistemological fusion of horizon(s). Only if we make this distinction and stress the ontological dimension of the shared, intersubjective horizon, do we realize that horizons are not purely epistemic.

17 We might wonder to what extent the seeming impossibility to fuse horizon(s) with the depressed person has not only to do with being exposed to sheer negativity. Ordinarily, we take the shared, intersubjective horizon we find ourselves in for granted and are thus less likely to become aware of it. Such an awareness, however, I think, seems to be required if we are trying to understand the depressed person's radical change of finding themselves in the world, precisely because for them this wider horizon has changed. Only if we become aware of this primordial horizon but also of the possibility of its at least partial breakdown, I propose, a non-depressed person might start to understand the experiences of the depressed person, or "radical Other," that significantly diverge from their own, or at least understand how significantly different the depressed person's experiences might be.

18 Although the here presented hermeneutic approach to empathy in relation to intersubjectivity does emphasize the dialogical nature of understanding the other and their experience, this should not be understood as proposing that there is no longer any difference between the first person and second person point of view on one's own experience. This point is crucial. For, although the depressed person is influenced by their interactions with the therapist, they might nonetheless experience themselves and the other in a certain way, allowing the possibility to misinterpret the other but also themselves (Ratcliffe 2017). The only possible way to counter this misinterpretation is, I believe, to emphasize what I would call the hermeneutic virtues of good will toward the other and epistemic humility toward oneself, both of which are likely to be established and fostered throughout a long-term interaction. For an account of the possibility of hermeneutic conversation and ethics, see Mehmel (2016).

References

Ahmed, S. (2010), *The Promise of Happiness*, Durham, NC: Duke University Press.

American Psychiatric Association (2013), *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Arlington, VA: American Psychiatric Publishing.

Bosco, F. M., Colle, L., De Fazio, S., Bono, A., Ruberti, S., and Tirassi, M. (2009), Th.o.m.a.s.: An Exploratory Assessment of Theory of Mind in Schizophrenic Subjects, *Consciousness and Cognition*, 18:306–319.

Brampton, S. (2008), *Shoot the Damn Dog: A Memoir of Depression*, London: Bloomsbury.

Carruthers, P. and Smith, P. K., eds. (1996), *Theories of Theories of Mind*, Cambridge: Cambridge University Press.

De Jaegher, H. (2013), Embodiment and Sense-Making in Autism, *Frontiers in Integrative Neuroscience*, 7:15.

Fuchs, T. (2010), Subjectivity and Intersubjectivity in Psychiatric Diagnosis, *Psychopathology*, 43:268–274.

Fuchs, T. (2013), Depression, Intercorporeality, and Interaffectivity, *Journal of Consciousness Studies*, 20:219–238.

Fuchs, T. (2015), Pathologies of Intersubjectivity in Autism and Schizophrenia, *Journal of Consciousness Studies*, 22:191–214.

Gadamer, H.-G. (1976), *Philosophical Hermeneutics*, D. E. Linge, trans. Berkeley: University of California Press.

Gadamer, H.-G. (1989), *Text and Interpretation*, pp. 21–51. In *Dialogue & Deconstruction: The Gadamer-Derrida Encounter*. D. Michelfelder and R. E. Palmer, eds. D. J. Schmidt and R. E. Palmer, trans. Albany: State University of New York Press.

Gadamer, H.-G., *Gesammelte Werke*, 7th edition. Tübingen: Mohr Siebeck (*GW*). Bd. 1: *Hermeneutik I: Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik*, 2010. Bd. 2: *Hermeneutik II: Wahrheit und Methode. Ergänzungen, Register*, 1993. Bd. 10: *Hermeneutik im Rückblick*, 1995.

Gadamer, H.-G. (2000), Subjectivity and intersubjectivity, subject and person, trans. Peter Adamson and David Vessey, *Continental Philosophy Review*, 33: 275–287.

Gadamer, H.-G. (2004), *Truth and Method*, 2nd edition. J. Weinsheimer and D. G. Marshall, trans. London: Continuum.

Gadamer, H.-G. (2006), The Incapacity for Conversation, D. Vessey and C.Blauwkamp, trans. *Continental Philosophy Review*, 39:351–359.

Gallagher, S. (2012), In Defense of Phenomenological Approaches to Social Cognition: Interacting with the Critics, *Review of Philosophy and Psychology*, 3:187–212.

Garrett, J. E. (1978), Hans-Georg Gadamer on "Fusion of Horizons," *Man and World*, 11:392–400.

Goldman, A. I. (2006), *Simulating Minds*, Oxford: Oxford University Press.

Gordon, R. M. (1986), Folk Psychology as Simulation, *Mind and Language*, 1:158–171.

Heal, J. (1996), Simulation and Cognitive Penetrability, *Mind and Language*, 11:44–67.

Hutto, D. D. (2013), Interpersonal Relating, pp. 240–257. In *The Oxford Handbook of Philosophy and Psychiatry*. K. W. M. Fulford, M. Davies, R. Gipps, G. Graham, J. Sadler, G. Stanghellini, and T. Thornton, eds. Oxford: Oxford University Press.

Karp, D. A. (1996), *Speaking of Sadness: Depression, Disconnection, and the Meanings of Illness*, Oxford: Oxford University Press.

Lévinas, E. (1969), *Totality and Infinity*, A. Lingis, trans. Pittsburgh, PA: Duquesne University Press.

Mehmel, C. A. (2016), Possibility of Hermeneutic Conversation and Ethics, *Theoria and Praxis*, 4. https://theoriandpraxis.journals.yorku.ca/index.php/theoriandpraxis/article/view/39749.

Ratcliffe, M. (2012), Phenomenology as a Form of Empathy, *Inquiry: An Interdisciplinary Journal of Philosophy*, 55:473–495.

Ratcliffe, M. (2014a), The Phenomenology of Depression and the Nature of Empathy, *Medicine, Health Care and Philosophy*, 17:269–280.

Ratcliffe, M. (2014b), The Structure of Interpersonal Experience, pp. 221–238. In *Phenomenology of Embodied Subjectivity*. D. Moran and R. T. Jensen, eds. Dordrecht: Springer.

Ratcliffe, M. (2017), Empathy and Psychiatric Illness, pp. 190–200. In *Routledge Handbook of Philosophy of Empathy*. H. Maiborn, ed. Abingdon: Routledge.

Sprong, M., Schothorst, P., Vos, E., Hox, J., and Van Engeland, H. (2007), Theory of Mind in Schizophrenia: Meta-Analysis, *British Journal of Psychiatry*, 191:5–13.

Stanghellini, G. (2013), Philosophical Resources for the Psychiatric Interview, pp. 321–358. In *The Oxford Handbook of Philosophy and Psychiatry*. K. W. M. Fulford, M. Davies, R. Gipps, G. Graham, J. Sadler, G. Stanghellini, and T. Thornton, eds. Oxford: Oxford University Press.

Stein, E. (2008), *Gesamtausgabe—Bd. 5: Zum Problem der Einführung*, Freiburg im Breisgau: Herder.
Stern, D. (1991), A Philosophy for the Embedded Analyst, *Contemporary Psychoanalysis*, 27:51–80.
Stich, S. and Nichols, S. (2003), *Mindreading*, Oxford: Clarendon Press.
Styron, W. (1990), *Darkness Visible*, London: Vintage.
Taylor-Aiken, A. (2011), Relatedness and Alienation in Interpersonal Understanding: A Phenomenological Account, PhD thesis, Durham University.
Thompson, T. (1996), *The Beast: A Journey through Depression*, Harmondsworth: Penguin Books.
Varga, S. (2015), *Naturalism, Interpretation and Mental Disorder*, Oxford: Oxford University Press.
Veith, J. (2015), *Gadamer and the Transmission of History*, Bloomington: Indiana University Press.
Vilhauer, M. (2009), Beyond the "Fusion of Horizons": Gadamer's Notion of Understanding as "Play", *Philosophy Today*, 53:359–364.
Vilhauer, M. (2010), *Gadamer's Ethics of Play*, Lanham, MD: Lexington Books.
Zahavi, D. (2011), Empathy and Direct Social Perception: A Phenomenological Proposal, *Review of Philosophy and Psychology*, 2:541–558.

Study Questions

1. *What is it like to have a mental disorder?*
2. *What is the best source of information about mental disorders: first person reports or third person reports?*
3. *Are first person accounts of mental disorders reliable?*
4. *How can the insights from phenomenology be useful in the clinical treatment of the patient?*

Part Four

Philosophy of Science and Psychiatry

9

Philosophy of Science, Psychiatric Classification, and the *DSM*

Jonathan Y. Tsou

Introduction

What is the purpose of classifying mental disorders? When should we regard a system of psychiatric classification as being successful? What kind of methodology should a system of psychiatric classification adopt? What are the relative benefits of descriptive as opposed to theoretical approaches to psychiatric classification? Should the *DSM* classify natural kinds? This chapter addresses these interrelated questions by critically examining the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*), which has been published regularly by the American Psychiatric Association since 1952 and is currently the most influential psychiatric classification system in the United States. While no definitive answers to these questions will be reached, my analysis aims to elucidate some of the philosophical considerations involved in assessing the *DSM*.

This chapter focuses on philosophical issues concerning the appropriate methodology for classifying mental disorders and whether the *DSM* should classify natural kinds. The chapter proceeds as follows. After briefly examining the history of the *DSM* and its adoption of a purely descriptive approach to classification in *DSM-III* (APA 1980), the explicit aims of the *DSM* (i.e., guiding treatment, facilitating research, and improving communication) are examined as a reference point for assessing the *DSM*. It is argued that the *DSM* is currently failing to meet its own aims, largely because of its failure to provide valid diagnostic categories (i.e., definitions that map onto real disorders in nature). As a potential solution to this problem, the prospects of a theoretical and causal approach to psychiatric classification are considered as an alternative to the *DSM*'s purely

descriptive approach. As a more fundamental and general philosophical issue concerning psychiatric classification, the question of whether the *DSM* is able to classify natural kinds is explored.

A brief history of the *DSM*

Before addressing philosophical questions regarding how mental disorders should be classified, it is important to examine some of the historical factors that have shaped the *DSM*. Prior to the publication of the *DSM*, the first standardized psychiatric classification manual published in the United States was the *Statistical Manual for the Use of Institutions for the Insane* (American Medico-Psychological Association and National Committee for Mental Hygiene 1918). The *Statistical Manual* was the definitive classification manual for mental disorders during the interwar years and it went through ten editions from 1918 to 1942. This manual was primarily intended for the collection of institutional data among inpatient mental hospitals and it implicitly assumed that mental disorders were biological disorders (Grob 1991b). After the Second World War, it was apparent that the *Statistical Manual* was insufficient to accommodate the types of patients that military psychiatrists treated during the war, the psychosocial problems that patients presented in outpatient settings, and the psychoanalytic assumptions that were increasingly becoming entrenched in early twentieth-century psychiatry (Horwitz 2015). These limitations of the *Statistical Manual* provided the impetus for a new system of psychiatric classification.

The first two editions of the *DSM* (APA 1952, 1968) classified mental disorders in a theoretical and etiological manner. Mental disorders were classified in a theoretical manner insofar as the disorders listed in these manuals were strongly informed by Sigmund Freud's psychoanalytic theory, which emphasized unconscious conflicts that contributed to mental illness, as well as Adolf Meyer's life-course approach to psychiatry, which emphasized how mental illnesses were reactions that individuals experienced in response to their environments (Horwitz 2015; cf. Cooper and Blashfield 2016). *DSM-I* and *DSM-II* classified disorders in an etiological manner insofar as the causes of disorders were often indicated, either in the name of the disorder or in the general class in which the disorder was classified. A fundamental classificatory distinction in the first two editions of the *DSM* was a distinction between disorders with biological causes and disorders with psychological causes (Tsou 2011). In *DSM-I*, this distinction was captured in the general classificatory distinction between "disorders caused

by or associated with impairment of brain tissue function" (APA 1952: 2) and "disorders of psychogenic origin without clearly defined physical cause or structural change in the brain" (APA 1952: 5). In *DSM-II*, this general etiological distinction was preserved under the new terms "organic brain syndromes" and "functional disorders" (APA 1968: 123).

The third edition of the *DSM* (APA 1980) marked a "revolution" in psychiatric classification in its explicit replacement of the theoretical and etiological approach adopted in *DSM*-I and *DSM*-III with an atheoretical and descriptive approach. In the introduction of *DSM*-III, the new approach to classification was described as follows:

> The approach taken in DSM-III is atheoretical with regard to etiology … except for those disorders for which this is well established … The major justification for the generally atheoretical approach … is that the inclusion of etiological theories would be an obstacle to use of the manual by clinicians of varying theoretical orientations … Because DSM-III is generally atheoretical with regard to etiology, it attempts to describe comprehensively what the manifestations of the mental disorders are, and only rarely attempts to account for how the disturbances come about … This approach can be said to be "descriptive" in that the definitions of the disorders generally consist of descriptions of the clinical features of the disorders. … at the lowest order of inference necessary to describe the characteristic features of the disorder. (APA 1980: 7)

This shift to a purely descriptive approach to classification was signified most prominently by the introduction of the now familiar "diagnostic criteria" in *DSM*-III, wherein individuals need to satisfy a set of necessary and sufficient criteria (mainly behavioral or otherwise observable criteria) to receive a *DSM* diagnosis. *DSM*-III and its successors also removed references to the causes of disorders and theoretical assumptions (viz., psychoanalytic assumptions) were excluded from the language of the manual. From a scientific point of view, *DSM*-III was heralded as a great success because its introduction of diagnostic criteria provided *reliable diagnostic categories* that different users could apply in a consistent manner. *DSM*-III was also successful in ensuring that the manual could be widely used by mental health professionals endorsing disparate theoretical assumptions. In this regard, it is not an accident that the *DSM* only became an influential and increasingly entrenched manual after *DSM*-III (Hacking 2013).

From a historical point of view, there are several factors responsible for the *DSM*'s shift to an atheoretical and descriptive approach to classification. On a general level, one of the most important factors was the declining status, in

the 1970s, of the psychoanalytic approach to psychiatry, which had informed the definitions of *DSM*-I and *DSM*-II. While psychoanalysis enjoyed a period of dominance in the United States from the postwar period to the early 1960s (Grob 1991b; Hale 1995; Wilson 1993), its scientific legitimacy was increasingly called into question after the 1960s (e.g., see Foucault 1965; Szasz 1961). Moreover, the introduction of the first antipsychotic drug in the 1950s led to a mass *deinstitutionalization* of mental patients insofar as treated patients no longer needed to be confined to asylums and could live in community settings (Grob 1991a, 1995). The deinstitutionalization of mental patients and increasing use of pharmacological drugs to treat mental disorders (e.g., schizophrenia, bipolar disorder, depression) in the 1960s and 1970s highlighted the inability of psychoanalytic methods to treat more serious disorders (Mayes and Horwitz 2005). In the 1970s, the *DSM* was also facing criticism from third-party insurance companies, who demanded clarity in psychiatric diagnoses and accountability for treatment outcomes (Horwitz 2002: ch. 3). By the mid-1970s, insurance companies such as Blue Cross and Aetna cut financial support for psychotherapy, citing lack of clarify and uniformity of diagnoses of mental disorders (Wilson 1993). As such, the psychoanalytically informed diagnostic categories of *DSM-II* were failing to meet third-party insurance companies' need for clear and standardized psychiatric diagnoses.

On a more specific level, the most decisive factor in the shift to a purely descriptive approach to classification was the "neo-Kraepelinian" outlook of the taskforce responsible for revising *DSM*-III. In 1974, Robert L. Spitzer was appointed to chair the *DSM*-III taskforce and he selected a number of like-minded individuals—with whom he had close professional allegiances—to join the taskforce (Blashfield 1984). In particular, Spitzer had close ties to a research group at Washington University in St. Louis who developed the "Feighner criteria": a set of diagnostic criteria for fifteen mental disorders, which effectively served as prototypes for *DSM*-III diagnostic categories (Feighner et al. 1972). Spitzer's taskforce shared a neo-Kraepelinian outlook on psychiatric classification insofar as they followed the tradition of psychiatry associated with Emil Kraepelin, who advocated a system of biological psychiatry that competed with Freudian psychoanalysis in the early twentieth century. Whereas psychoanalysts assumed that psychological mechanisms play a key role in mental illness and the boundary between normality and abnormality is continuous (ranging from neuroses to psychoses), the neo-Kraepelinians assumed that mental illnesses are *discrete biological diseases* and there is a clear boundary between the normal and the sick (Blashfield 1984; Klerman 1978; Wilson

1993). From the neo-Kraepelinian perspective, it was crucial to distinguish the mentally ill from the well with *reliable and valid diagnostic categories*. The desideratum for empirically validated diagnostic categories motived the neo-Kraepelinian belief that mental disorders ought to be classified on the basis of their observable symptoms, rather than on the basis of speculative inferences regarding their causes. This methodological stance constituted the foundation for the purely descriptive approach adopted in *DSM*-III. Since the publication of *DSM*-III, the purely descriptive, neo-Kraepelinian methodology has remained the status quo in subsequent editions of the *DSM*.

Aims of the *DSM*

In the introduction to the fourth edition of the *DSM*, one finds one of the most explicit and clearest statements of the goals of the *DSM*:

> The utility and credibility of DSM-IV require that it focus on its clinical, research, and educational purposes and be supported by an extensive empirical foundation. *Our highest priority has been to provide a helpful guide to clinical practice. … An additional goal was to facilitate research and improve communication among clinicians and researchers*. We were also mindful of the use of the DSM-IV for improving the collection of clinical information and as an educational tool for teaching psychopathology. (APA 2000: xxiii, emphasis added; cf. APA 2013: xli)

As suggested here, the main aims of the *DSM* are to: (1) provide a manual that can guide treatment decisions, (2) facilitate research, and (3) improve communication among mental health professionals. These explicitly cited aims of the *DSM* present a reasonable starting point for evaluating the *DSM* as a psychiatric classification manual.

How well does the *DSM* do in meeting these three aims of the *DSM*? As indicated above, the *DSM* states that its primary goal is to provide a helpful guide in clinical and treatment contexts. Unfortunately, there is evidence that the *DSM* does not fare very well on this particular goal. While *DSM*-III was successful in providing reliable diagnostic categories that can be applied uniformly by clinicians working in different treatment contexts, there is little evidence that the *DSM*'s fine-grained diagnostic categories are actually useful for clinicians in guiding treatment decisions. In clinical contexts, a *DSM* diagnosis is necessary for patients seeking treatment to be reimbursed for services. Research has indicated that clinicians using the *DSM* frequently deliberately give inaccurate diagnoses to patients; that is, giving a less severe diagnosis to avoid the stigmatizing

effects of diagnostic labels or providing a diagnosis so that clients can receive reimbursement for treatment services (Kirk and Kutchins 1988, 1992: ch. 9). Others have discussed the "sociological ambivalence" experienced by clinicians (Brown 1987; Whooley 2010), in which the demands placed on clinicians by the *DSM* conflict with their desire to exercise individual professional autonomy in helping individuals seeking treatment. This leads to situations where clinicians will intentionally fudge *DSM* codes provided on insurance forms, or negotiate diagnoses with clients (Whooley 2010). These kinds of practices fall well short of the diagnosis-treatment model implicitly assumed in the *DSM*, wherein clinicians diagnose patients with a *DSM* category and provide a treatment plan on the basis of the diagnosis. In addition to intentional misdiagnoses, a well-known difficulty with the *DSM* in practice is the high incidence of comorbid diagnoses (Kessler et al. 2005), which undermines the entire rationale of the diagnosis-treatment model. In actual clinical practices, one of the most useful *DSM* tools is the "not otherwise specified" (NOS) code, which indicates that a patient does not fit neatly into any of the *DSM*'s diagnostic categories, but allows individuals to receive a *DSM* diagnosis for the purposes of medical insurance (Hyman 2010). Hence, in clinical contexts, the *DSM* appears to function as an administrative constraint, rather than as a helpful guide for clinical practice and treatment decisions.

There are also reasons for thinking that the *DSM* is not particularly useful—and perhaps an impediment—for facilitating research on mental disorders. The *DSM* and its diagnostic categories currently guide research by providing operational definitions of mental disorders, which researchers utilize to select homogenous populations of patients to study. While this strategy is useful for ensuring that researchers working in different locales are studying disorders in a uniform manner, a fundamental flaw with this research strategy stems from the fact that while *DSM* categories are reliable, they lack validity in the sense that we have no assurance that *DSM* diagnostic categories map onto discrete mental disorders in reality (Kendell 1989; Kendell and Jablensky 2003). This presents a situation where the *DSM* may function to *reify* operationally defined categories that do not have any basis in reality (Hyman 2010). Frustration with the lack of validity of *DSM* categories has led the National Institute of Mental Health to develop the Research Domain Criteria (RDoC), which is an alternative psychiatric classification system intended primarily for research on psychopathology. One of the motivations of the RDoC was to counter the problem that *DSM* categories do not correspond well to emerging findings from genetics, behavioral science, and neuroscience (Cuthbert and Insel 2013; Hyman 2007; Sanislow et al. 2010).

The RDoC is premised on the idea that shifting research away from the *DSM*'s operational definitions will provide a more fruitful approach for identifying valid diagnostic categories (or constructs). Rather than studying psychopathology by focusing on the specific symptom-based definitions of mental disorders provided by the *DSM*, the RDoC attempts to direct research on psychopathology toward more general constructs (e.g., potential threat, reward learning, cognitive control) that are more likely to be explicable at lower levels of analysis (e.g., genes, neural circuits, physiology), and hence serve as more promising targets of validation than *DSM* definitions (Bluhm 2017; Tabb 2015).

While the *DSM* is not highly successful at guiding clinical treatment or facilitating research, it does undoubtedly improve communication among different mental health professionals. The *DSM* achieves this aim through its provision of standardized operational definitions of mental disorders, which can be used in a variety of different contexts, including contexts related to treatment and research. Relatedly, *DSM* definitions of mental disorders can be widely used by a variety of different users since *DSM* definitions make no strong theoretical or metaphysical assumptions about the nature or causes of mental disorders. Hence, the *DSM* helps to standardize and coordinate communication about psychopathology through its provision of a common language for talking about mental disorders. While this benefit of the *DSM* should not be underestimated, this virtue needs to be balanced against the lack of validity of *DSM* definitions. If *DSM* categories are not successfully individuating real disorders in nature, then the benefit of communicating in the language of *DSM* categories may not be a substantial benefit, but a drawback.

Descriptive vs. theoretical approaches

Given the explicitly stated aims of the *DSM* (i.e., guiding treatment, facilitating research, and improving communication), what is an effective methodology for achieving these goals? As indicated above, the purely descriptive, neo-Kraepelinian methodology adopted by the *DSM* since *DSM*-III has failed to satisfactorily meet the *DSM*'s goals, with exception of improving communication among different mental health professionals. The failure of the *DSM* to provide a useful manual for informing treatment decisions and for facilitating research stems largely from the inability of the *DSM* to identify valid diagnostic categories.

Several philosophers of science have argued in favor of a theoretical and causal approach to psychiatric classification. In his classic analysis of psychiatric

taxonomy, Carl Hempel (1965) argued that a good taxonomic system in psychiatry must not only reliably describe its objects of classification, but it should also facilitate the simultaneous formulation of scientific theories and laws:

> [G]ranting some oversimplification, the development of a scientific discipline may often be said to proceed from an initial "natural history" stage, which primarily seeks to describe the phenomena under study and to establish simple empirical generalizations concerning them, to subsequent more and more "theoretical" stages, in which increasing emphasis is placed upon the attainment of *comprehensive theoretical accounts of the empirical subject matter under investigation*. (Hempel 1965: 140, emphasis added)

According to Hempel, the classificatory concepts of taxonomic systems serve two main epistemic functions: (1) description and (2) explanation. With respect to the descriptive function of classificatory concepts, Hempel recommends the usage of operational definitions to objectively define its objects of classification. While Hempel would have approved of the *DSM*'s introduction of operational definitions in *DSM*-III, he would also maintain that for a taxonomic system to achieve a truly scientific status, it must move beyond this descriptive stage into increasingly theoretical stages. With respect to the explanatory function of classificatory concepts, Hempel recommends that concepts should be located within a broader theoretical framework that provides a systematic way for organizing the subclasses of that system. For this end, Hempel (1965) recommends the formulation of general laws and theoretical principles:

> [T]o be scientifically useful *a concept must lend itself to the formulation of general laws or theoretical principles which reflect uniformities in the subject matter under study, and which thus provides a basis for explanation, prediction, and generally scientific understanding*. (146, emphasis added)

Hempel argues that psychiatric classification systems ought to reach a theoretical stage because this will ensure that its classificatory concepts are explanatory and predictive. From Hempel's perspective, if a classification system is to provide projectable inferences (i.e., reliable predictions) about its objects of classification (e.g., inferences regarding the prognosis of a disorder, or which treatment interventions might be successful), its concepts should be associated with scientific theories or laws (e.g., psychoanalytic or physiological theories), which make explicit hypotheses regarding the causes of the disorders.

More recently, Dominic Murphy (2006) and Jonathan Tsou (2015, 2016) have argued that a theoretical and causal approach to psychiatric classification

based on biological psychiatry would offer a more promising method for classifying mental disorders than the purely descriptive approach currently favored by the *DSM*. Murphy (2006) argues that the *DSM*'s purely descriptive system of classification is incoherent in its application of the medical model. While the *DSM* implicitly assumes that mental disorders are distinct (disease) entities that reflect underlying causal differences, it prohibits reference to the causal processes and dysfunctions that distinguish them. Hence, the *DSM*'s purely descriptive approach to classification provides an impoverished (or incoherent) method for individuating mental disorders that are assumed to be causally distinctive diseases. Murphy also maintains that the *DSM*'s symptom-based approach to classification fails to distinguish heterogeneous conditions since *DSM* categories group together different symptom profiles as manifestations of the same disorder. In place of the *DSM*'s neo-Kraepelinian approach, Murphy advocates a theoretical and causal approach to classification that supplements the descriptive categories of the *DSM* with causal information concerning the determinants of various symptoms. As Murphy (2006) puts it, "[by] articulating causal explanations … , the hope is that we will eventually be able [to] discriminate more finely among conditions that are currently lumped together by *DSM*'s fuzzy categories, by distinguishing syndromes on the basis of etiology and pathology" (349–350). Similarly, Tsou (2015, 2016) argues that one of the main deficits of the *DSM*'s purely descriptive approach is that it fails to facilitate a testable and revisable system of classification and the *DSM* has functioned to reify and entrench artificial kinds of mental disorders (e.g., histrionic personality disorder). If the *DSM* aims to classify naturally occurring kinds of mental disorders (i.e., natural kinds), Tsou (2015, 2016) argues that diagnostic categories should integrate causal information such that the operational criteria that are used to define mental disorders are associated (or hypothesized to be associated) with identifiable biological mechanisms (e.g., neurobiological mechanisms). Tsou argues that, compared to the *DSM*'s purely descriptive approach, a theoretical and causal approach to classification would provide a superior and more transparent methodology for identifying valid diagnostic categories and meeting the *DSM*'s goals (especially the goals of guiding treatment and facilitating research).

Disenchantment with the failure of the *DSM* to identify valid diagnostic categories has also led some mental health researchers to advocate a theoretical approach to psychiatric classification. Prior to the publication of *DSM*-5, some members of the *DSM*-5 taskforce—led by David J. Kupfer—indicated that *DSM*-

5 would mark a shift toward a theoretical approach to classification informed by sciences such as genetics, neuroscience, and the cognitive sciences (e.g., see Hyman 2007; Kupfer, First, and Regier 2002; Kupfer, Kuhl, and Regier 2013; Kupfer and Regier 2011; Regier 2008; Regier et al. 2009). For example, Kupfer, First, and Regier (2002) discuss the "[n]eed to explore the possibility of fundamental changes in the neo-Kraepelinian diagnostic paradigm" (xviii) and the potential of the *DSM*-5 to usher a paradigm-shift toward a theoretical approach to psychiatric classification:

> Concerns have … been raised that researchers' slavish adoption of DSM-IV definitions may have hindered research in the etiology of mental disorders. Few question the value of having a well-described, well-operationalized, and universally accepted diagnostic system to facilitate diagnostic comparisons across studies and to improve diagnostic reliability. However, reification of DSM-IV entities, to the point that they are considered to be equivalent to diseases, is more likely to obscure than to elucidate research findings.
>
> All these limitations in the current [neo-Kraepelinian] diagnostic paradigm suggest that research exclusively focused on refining the DSM-defined syndromes may never be successful in uncovering their underlying etiologies. For that to happen, an as yet unknown paradigm shift may need to occur. Therefore, another important goal … is to transcend the limitations of the current DSM paradigm and to encourage a research agenda that goes beyond our current ways of thinking to attempt to integrate information from a wide variety of sources and technologies. (xix)

The eventual publication of *DSM*-5 (APA 2013) was disappointing for those hoping for revisions of paradigm-shifting proportions; the main discernible methodological difference in *DSM*-5 compared to its predecessors is the greater use of dimensional measures.

As discussed earlier, the repeated failure of the *DSM* to identify valid diagnostic categories also spurned the National Institute of Mental Health to introduce the RDoC project, which supports an explicitly theoretical approach to psychiatric classification intended primarily for facilitating research on psychopathology (Casey et al. 2013; Cuthbert 2014; Cuthbert and Insel 2013; Insel et al. 2010). A large part of RDoC project is the RDoC matrix (Morris and Cuthbert 2012: 31), which seeks to organize and facilitate psychopathology research by distinguishing seven units of analysis (viz., genes, molecules, cells, circuits, physiology, behavior, self-report, and paradigms) and five research domains (viz., negative valence emotional systems, positive valence emotional systems, cognitive systems, social processes, and arousal/

regulatory systems). The RDoC matrix offers an alternative classification system to the *DSM* for research purposes, and it is premised on the idea that research exclusively framed in terms of *DSM* categories are inadequate for formulating valid diagnostic categories (Cuthbert and Insel 2013; Insel et al. 2010).

Despite these various arguments for a theoretical and causal approach to psychiatric classification by philosophers of science and practicing mental health researchers, the *DSM* has upheld its purely descriptive, neo-Kraepelinian approach to classification. Proponents of the *DSM*'s purely descriptive approach have emphasized the need to be conservative in revising the *DSM* (Frances 2009; Frances and Widiger 2012); while others have stressed that the *DSM* is not prepared for full-scale revolutionary changes to its methodology (Kendler and First 2010).

Can the *DSM* classify natural kinds?

Some commentators on psychiatric classification have argued that the proper objects or targets of psychiatric classification are natural kinds; that is, classes of mental disorders that have a shared natural basis (e.g., Tsou 2016). In philosophy of science, natural kinds (e.g., H_2O, electrons, quarks) are understood as classes or groupings that accurately correspond to the natural structure of the world, whereas artificial kinds (e.g., chairs, books, games) are classes that reflect the interests and actions of humans (Bird and Tobin 2017). Hence, natural kinds are thought to be naturally occurring ("real") classes that are the proper objects of scientific study; artificial kinds are arbitrary or invented classes that may serve some social function, but do not have a natural basis. Several philosophers of science have offered reasons for thinking that the *DSM* is unable to classify natural kinds.

While there is no agreement on what natural kinds in psychiatry are (e.g., see Beebee and Sabbarton-Leary 2010; Cooper 2005, 2007; Haslam 2002; Kendler, Zachar, and Craver 2011; Kincaid and Sullivan 2014; Samuels 2009; Tsou 2013, 2016; Zachar 2000), mental disorders that are natural kinds are generally conceptualized as classes of abnormal behavior that have a clear natural (as opposed to artificial) basis. With respect to human science classifications, some classifications (e.g., "introversion," "Hodgkin's Lymphoma," "Down Syndrome") refer to natural kinds insofar as these classes can be defined with reference to a set of natural (i.e., biological) properties; other classes (e.g.,

"widows," "liberals," "policemen") refer to artificial kinds insofar as these classes are defined exclusively in terms of a set of conventionally defined (i.e., social) properties. While most mental disorders listed in the *DSM* include conventional aspects, it is reasonable to assume that some of them are "more natural" insofar as these classes—to a large extent—can be individuated with reference to a set of natural properties. For example, forms of schizophrenia that include psychotic elements (i.e., auditory hallucinations and delusions) can be defined with reference to identifiable dysfunctional neurobiological processes, such as excessive dopamine activity in the mesolimbic pathway (see Tsou 2012).

An implicit goal of the *DSM* since the publication of *DSM*-III is that the manual should classify mental disorders that are natural kinds. Prior to the publication of *DSM*-III, the authors of *DSM*-III planned to include a statement in the manual indicating that the mental disorders listed in the *DSM* are "a subset of medical conditions" (Spitzer and Endicott 1978; Spitzer, Sheehy, and Endicott 1977), which suggests that the *DSM* aims to classify natural kinds (viz., naturally occurring disease entities). This statement, however, was ultimately not included in *DSM*-III due to protests by psychologists, social workers, and counselors, who regarded this statement as a declaration that psychiatrists with medical training were solely responsible for the treatment of mental disorders (Mayes and Horwitz 2005). The general definition of mental disorder presented in *DSM*-5 also indicates that the *DSM* aims to classify natural kinds: "A mental disorder is a syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behavior that *reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning*" (APA 2013: 20, emphasis added; cf. APA 2000: xxxi). This definition indicates that the mental disorders classified by the *DSM* are natural kinds insofar as these disorders are caused by a dysfunction in some naturalistic (i.e., psychological, biological, or developmental) process. Significantly, classes of abnormal behavior that merely deviate from social norms—as opposed to naturalistic norms—are explicitly excluded from this definition of mental disorder (APA 2013: 20).

Ian Hacking has offered an influential argument for thinking that the *DSM* cannot classify natural kinds. Hacking (1999) argues that—in contrast to the stable objects of classification in the natural sciences (e.g., electrons, quarks)—the objects of classification in psychiatry (and the human sciences more generally) are inherently unstable because of the "looping effects" of psychiatric classifications. Looping effects are social feedback effects of human science

classifications, wherein individuals who are classified change in response to how they are classified, for example, children who are diagnosed with attention deficit hyperactivity disorder (ADHD) act in accordance with the stereotypes associated with this label (Hacking 1995). According to Hacking, looping effects are specific to human science classifications because the objects of classification in the human sciences (unlike the objects of classification in the natural sciences) are aware of how they are classified and will change in response to how they are classified. Hacking (2007) writes:

> We think of … kinds of people [e.g., obese people or impoverished people] … as definite classes defined by definite properties. … But it is not quite like that. They are moving targets because our investigations interact with the targets themselves, and change them. And since they are changed, they are not quite the same kind of people as before. The target has moved. That is the looping effect. (293)

According to Hacking, the objects of classification in psychiatry (e.g., depressed people, autistic children) are constantly reacting and changing in response to how they are classified and the classifications for these kinds of people will need to be constantly revised to accommodate such changes, which in turn will cause further changes in the people so classified. This captures the sense in which Hacking wants to suggest that the objects of classification (i.e., kinds of people) in psychiatry and the human science are "moving targets" because of looping effects.

Several commentators have contested Hacking's conclusion that the objects of classification in the human sciences are not natural kinds because of looping effects (Cooper 2004; Tsou 2007; cf. Tekin 2014, 2016). Rachel Cooper (2004) argues that Hacking fails to establish that the objects classified in the human sciences are sufficiently different from the objects of classification in the natural sciences. While Cooper grants the existence of looping effects, she argues that the changes that looping effects produce for objects of classification in the human sciences are not of greater metaphysical significance than changes that other kinds of feedback effects produce for some objects of classification in the natural sciences. For instance, due of classificatory practices, the phenotypic traits of domesticated animals can change because of the ways that animals are selectively bred, or the use of antibiotics can cause mutations in bacteria (Bogen 1988; Khalidi 2010). Hence, Cooper concludes that while looping effects may be unique to human science classifications, this should not lead us to think that the objects classified in the human sciences are significantly

different from the objects classified in the natural sciences with respect to the stability of these respective kinds. From a different perspective, Tsou (2007) argues that Hacking's inference from the existence of looping effects to the instability of objects of classification in psychiatry is illegitimate. Tsou grants the putative existence of looping effects, but denies that the mere presence of looping effects will inevitably render objects of classification unstable in the way suggested by Hacking. For example, while the experiences and behavior of an individual who is diagnosed with depression will invariably be altered by the way that he or she is classified, these changes will not necessarily require revisions in the depression classification or changes in what it means to be clinically depressed. So long as depression can—to a large extent—be identified with a stable set of biological properties (e.g., deficient activity of serotonin) that determine a stereotypical set of abnormal emotional states (e.g., feelings of sadness, loss of pleasure), depression is a natural kind that remains stable in spite of the looping effects of the depression classification (Tsou 2013).

Cooper (2005) has offered a somewhat different argument for thinking that the *DSM* is unable to classify natural kinds. Cooper points out that the *DSM* does not exist in a historical and social vacuum and that the *DSM* has been strongly shaped by social forces. From this perspective, Cooper argues that the *DSM*'s goal of classifying natural kinds has been impeded by social and financial factors from the pharmaceutical and medical industries. With respect to the pharmaceutical industry, Cooper brings attention to the way that the *DSM* has been shaped by the marketing of particular drug treatments. For instance, some mental disorders (e.g., panic disorder, social phobia) have been substantiated through marketing campaigns aimed to show that they can be successfully treated by particular drugs. Similarly, she suggests that niche marketing of certain mental disorders (e.g., attention deficit hyperactivity disorder) by pharmaceutical companies can function to raise the prominence (and prevalence) of a condition, thereby helping to establish its legitimacy as a mental disorder. Cooper contends that the medical insurance industry presents an even stronger impediment to classifying natural kinds. The pervasive influence of this industry arises from the role that the *DSM* plays in medical insurance reimbursement. By the late 1960s, medical insurance carriers in the United States required a *DSM* diagnosis for patients to receive reimbursement for mental health treatment. According to Cooper, this created new pressures on the *DSM*, leading to new disorders (e.g., posttraumatic stress disorder) being included in *DSM*-III as the result of lobbying from certain groups to

receive reimbursement for treatment (see Young 1995). Because of such social constraints on the *DSM*, Cooper concludes that the *DSM*'s goal of classifying natural kinds is unlikely to be achieved in practice.

Conclusion: Assessing the *DSM*

This chapter focused on two philosophical issues concerning the classification of mental disorders: (1) whether a theoretical and causal approach to classification would be superior to the purely descriptive approach currently favored by the *DSM* and (2) whether the *DSM* is capable of classifying natural kinds. With respect to the first issue, there are good philosophical reasons for thinking that the *DSM* should shift away from its purely descriptive, neo-Kraepelinian stance; however, there are also compelling pragmatic reasons for maintaining the *DSM*'s purely descriptive methodology. With respect to the second issue, the *DSM*'s ideal of classifying natural kinds (and formulating valid diagnostic categories) appears to be a reasonable philosophical ideal for psychiatric classification; however, there are philosophical and pragmatic reasons for thinking that this ideal may be unattainable in practice.

Returning to the more general issue of assessing the *DSM*, the *DSM*'s explicit goals of providing a manual that can guide treatment, facilitate research, and improve communication appear to be defensible goals. However, as argued in this chapter, the *DSM* is currently failing to substantially achieve these various goals, which can largely be attributed to its failure to provide valid diagnostic categories. This raises the question of whether the aims of the *DSM* ought to be defined more narrowly than they currently are, for example, to provide a manual whose sole purpose is to improve communication by providing a common reference point that allows for the exchange of information and integration of explanations among various mental health professionals (Kutschenko 2011). Perhaps one of the largest difficulties with the *DSM* is that it has formulated its aims too broadly and its hegemonic status has prevented the serious consideration of alternative, more specifically oriented, psychiatric classification systems, for example, classification systems formulated exclusively for treatment (Tsou 2015). As discussed in this chapter, the current shape that the *DSM* has taken—including its aims—has arisen out of contingent historical factors. Hence, it may be time to reconsider the goals of the *DSM* in light of our current historical situation and the specific problems that it poses.

Acknowledgments

This chapter draws on material developed in Tsou (2011, 2015, 2016). I am grateful to Robyn Bluhm and Şerife Tekin for helpful comments and suggestions.

References

American Medico-Psychological Association and National Committee for Mental Hygiene (1918), *Statistical Manual for the Use of Institutions of the Insane*, New York: American Medico-Psychological Association.

APA (1952), *Diagnostic and Statistical Manual: Mental Disorders*, Washington, DC: American Psychiatric Association.

APA (1968), *Diagnostic and Statistical Manual of Mental Disorders, Second Edition: DSM-II*, Washington, DC: American Psychiatric Association.

APA (1980), *Diagnostic and Statistical Manual of Mental Disorders, Third Edition: DSM-III*, Washington, DC: American Psychiatric Association.

APA (2000), *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition Text Revision: DSM-IV-TR*, Washington, DC: American Psychiatric Association.

APA (2013), *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition: DSM-5*, Washington, DC: American Psychiatric Association.

Beebee, H. and Sabbarton-Leary, N. (2010), Are Psychiatric Kinds "Real"?, *European Journal of Analytic Philosophy*, 6:11–27.

Bird, A. and Tobin, E. (2017), Natural Kinds. In *The Stanford Encyclopedia of Philosophy*. E. N. Zalta, ed. Spring 2017 edition. URL= https://plato.stanford.edu/archives/spr2017/entries/natural-kinds/

Blashfield, R. K. (1984), *The Classification of Psychopathology: Neo-Kraepelinian and Quantitative Approaches*, New York: Plenum Press.

Bluhm, R. (2017), Evidence-Based Medicine, Biological Psychiatry, and the Role of Science in Medicine, pp. 37–58. In *Extraordinary Science and Psychiatry: Responses to the Crisis in Mental Health Research*. J. Poland and Ş. Tekin, eds. Cambridge, MA: MIT Press.

Bogen, J. (1988), Comments on "The Sociology of Knowledge about Child Abuse," *Noûs*, 22:65–66.

Brown, P. (1987), Diagnostic Conflict and Contradiction in Psychiatry, *Journal of Health and Social Behavior*, 28:37–50.

Casey, B. J., Craddock, N., Cuthbert, B. N., Hyman, S. E., Lee, F. S., and Ressler, K. J. (2013), DSM-5 and RDoC: Progress in Psychiatric Research?, *Nature Reviews Neuroscience*, 14:810–814.

Cooper, R. (2004), Why Hacking Is Wrong about Human Kinds, *British Journal for the Philosophy of Science*, 55:73–85.

Cooper, R. (2005), *Classifying Madness: A Philosophical Examination of the Diagnostic and Statistical Manual of Mental Disorders*, Dordrecht: Springer.

Cooper, R. (2007), *Psychiatry and Philosophy of Science*, Montreal and Kingston: McGill-Queen's University Press.

Cooper, R. and Blashfield, R. K. (2016), Reevaluating the DSM-I, *Psychological Medicine*, 46:449–456.

Cuthbert, B. N. (2014), The RDoC Framework: Facilitating Transition from ICD/DSM to Dimensional Approaches That Integrate Neuroscience and Psychopathology, *World Psychiatry*, 13:28–35.

Cuthbert, B. N. and Insel, T. R. (2013), Toward the Future of Psychiatric Diagnosis: The Seven Pillars of RDoC, *BMC Medicine*, 11:126.URL= http://www.biomedcentral.com/1741-7015/11/126.

Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G., and Munoz, R. (1972), Diagnostic Criteria for Use in Psychiatric Research, *Archives of General Psychiatry*, 26:57–63.

Foucault, M. (1965), *Madness and Civilization: A History of Insanity in the Age of Reason*. R. Howard, trans. New York: Pantheon.

Frances, A. J. (2009), A Warning Sign on the Road to DSM-V: Beware of Its Unintended Consequences, *Psychiatry Times*, 26:1–4.

Frances, A. J. and Widiger, T. (2012), Psychiatric Diagnosis: Lessons from the DSM-IV Past and Cautions for the DSM Future, *Annual Review of Clinical Psychology*, 8:109–130.

Grob, G. N. (1991a), *From Asylums to Community: Mental Health Policy in Modern America*, Princeton, NJ: Princeton University Press.

Grob, G. N. (1991b), Origins of *DSM-I*: A Study of Appearance and Reality, *American Journal of Psychiatry*, 148:421–431.

Grob, G. N. (1995), The Paradox of Deinstitutionalization, *Society*, 32:51–59.

Hacking, I. (1995), The Looping Effects of Human Kinds, pp. 351–383. In *Causal Cognition: A Multidisciplinary Debate*. D. Sperber, D. Premack, and A. J. Premack, eds. Oxford: Clarendon Press.

Hacking, I. (1999), Madness: Biological or Constructed?, pp. 100–124. In *The Social Construction of What?*. I. Hacking, ed. Cambridge, MA: Harvard University Press.

Hacking, I. (2007), Kinds of People: Moving Targets, *Proceedings of the British Academy*, 151:285–318.

Hacking, I. (2013), Lost in the Forest, *London Review of Books*, 35:7–8.

Hale, N. G. (1995), *The Rise and Crisis of Psychoanalysis in America: Freud and the Americans, 1917–1985*, Oxford: Oxford University Press.

Haslam, N. (2002), Kinds of Kinds: A Conceptual Taxonomy of Psychiatric Categories, *Philosophy, Psychiatry, & Psychology*, 9:203–217.

Hempel, C. G. (1965), Fundamentals of Taxonomy, pp. 137–154. In *Aspects of Scientific Explanation: And Other Essays in the Philosophy of Science*. C. G. Hempel, ed. New York: The Free Press.

Horwitz, A. V. (2002), *Creating Mental Illness*, Chicago, IL: University of Chicago Press.

Horwitz, A. V. (2015), *DSM-I* and *DSM-II*, pp. 1–6. In *The Encyclopedia of Clinical Psychology*. R. I. Cautin and S. O. Lilienfeld, eds. URL= http://onlinelibrary.wiley.com/doi/10.1002/9781118625392.wbecp012/pdf.

Hyman, S. E. (2007), Can Neuroscience Be Integrated into the DSM-V?, *Nature Reviews Neuroscience*, 8:725–732.

Hyman, S. E. (2010), The Diagnosis of Mental Disorders: The Problem of Reification, *Annual Review of Clinical Psychology*, 6:155–179.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., and Wang, P. (2010), Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders, *American Journal of Psychiatry*, 167:748–751.

Kendell, R. E. (1989), Clinical Validity, *Psychological Medicine*, 19:45–55.

Kendell, R. and Jablensky, A. (2003), Distinguishing between the Validity and Utility of Psychiatric Diagnoses, *American Journal of Psychiatry*, 160:4–12.

Kendler, K. S. and First, M. B. (2010), Alternative Futures for the DSM Revision Process: Iteration *v*. Paradigm Shift, *British Journal of Psychiatry*, 197:263–265.

Kendler, K. S., Zachar, P., and Craver, C. (2011), What Kinds of Things Are Psychiatric Disorders?, *Psychological Medicine*, 41:1143–1150.

Kessler, R. C., Chiu, W. T., Demler, O., and Walters, E. E. (2005), Prevalence, Severity, and Comorbidity of 12-Month DSM-IV Disorders in the National Comorbidity Survey Replication, *Archives of General Psychiatry*, 62:617–627.

Khalidi, M. A. (2010), Interactive Kinds, *British Journal for the Philosophy of Science*, 61:335–360.

Kincaid, H. and Sullivan, J. A., eds. (2014), *Classifying Psychopathology: Mental Kinds and Natural Kinds*, Cambridge, MA: MIT Press.

Kirk, S. A. and Kutchins, H. (1988), Deliberate Misdiagnosis in Mental Health Practice, *Social Service Review*, 62:225–237.

Kirk, S. A. and Kutchins, H. (1992), *The Selling of DSM: The Rhetoric of Science in Psychiatry*, New York: Aldine de Gruyter.

Klerman, G. L. (1978), The Evolution of a Scientific Nosology, pp. 99–121. In *Schizophrenia: Science and Practice*. J. C. Shershow, ed. Cambridge, MA: Harvard University Press.

Kupfer, D. J., First, M. B., and Regier, D. A. (2002), Introduction, pp. xv–xxiii. In *A Research Agenda for DSM-V*, Washington, DC: American Psychiatric Association.

Kupfer, D. J., Kuhl, E. A., and Regier, D. A. (2013), *DSM-5*—The Future Arrived, *Journal of the American Medical Association*, 309:1691–1692.

Kupfer, D. J. and Regier, D. A. (2011), Neuroscience, Clinical Evidence, and the Future of Psychiatric Classification in *DSM-5*, *American Journal of Psychiatry*, 168:672–674.

Kutschenko, L. K. (2011), How to Make Sense of Broadly Applied Medical Classification Systems: Introducing Epistemic Hubs, *History and Philosophy of the Life Sciences*, 33:583–602.

Mayes, R. and Horwitz, A. V. (2005), DSM-III and the Revolution in the Classification of Mental Illness, *Journal of the Behavioral Sciences*, 41:32–52.

Morris, S. E. and Cuthbert, B. N. (2012), Research Domain Criteria: Cognitive Systems, Neural Circuits, and Dimensions of Behavior, *Dialogues in Clinical Neuroscience*, 14:29–37.

Murphy, D. (2006), *Psychiatry in the Scientific Image*, Cambridge, MA: MIT Press.

Regier, D. A. (2008), Forward: Dimensional Approaches to Psychiatric Classification, pp. xvii–xxiii. In *Dimensional Approaches in Diagnostic Classification: Refining the Research Agenda for DSM-V*. J. E. Helzer, H. C. Kraemer, R. F. Krueger, W. Hans-Ulrich, P. J. Sirovatka, and D. A. Regier, eds. Washington, DC: American Psychiatric Association.

Regier, D. A., Narrow, W. E., Kuhl, E. A., and Kupfer, D. J. (2009), The Conceptual Development of DSM-V, *American Journal of Psychiatry*, 166:645–650.

Samuels, R. (2009), Delusion as a Natural Kind, pp. 49–82. In *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*. M. R. Broome and L. Bortolotti, eds. Oxford: Oxford University Press.

Sanislow, C. A., Pine, D. S., Quinn, K. J., Kozak, M. J., Garvey, M. A., Heinssen, R. K., Wang, P. S., and Cuthbert, B. N. (2010), Developing Constructs for Psychopathology Research: Research Domain Criteria, *Journal of Abnormal Psychology*, 119:631–639.

Spitzer, R. L. and Endicott, J. (1978), Medical and Mental Disorder: Proposed Definition and Criteria, pp. 15–39. In *Critical Issues in Psychiatric Diagnosis*. R. L. Spitzer and D. F. Klein, eds. New York: Raven Press.

Spitzer, R. L., Sheehy, M., and Endicott, J. (1977), DSM-III: Guiding Principles, pp. 1–24. In *Psychiatric Diagnoses*. V. M. Rakoff and H. B. Kedward, eds. New York: Brunner/ Mazel.

Szasz, T. S. (1961), *The Myth of Mental Illness: Foundations of a Theory of Personal Conduct*, New York: Paul B. Hoeber.

Tabb, K. (2015), Psychiatric Progress and the Assumption of Diagnostic Discrimination, *Philosophy of Science*, 82:1047–1058.

Tekin, Ş. (2014), The Missing Self in Hacking's Looping Effects, pp. 227–256. In *Classifying Psychopathology: Mental Kinds and Natural Kinds*. H. Kincaid and J. A. Sullivan, eds. Cambridge, MA: MIT Press.

Tekin, Ş. (2016), Are Mental Disorders Natural Kinds? A Plea for a New Approach to Intervention in Psychiatry, *Philosophy, Psychiatry, & Psychology*, 23:147–163.

Tsou, J. Y. (2007), Hacking on the Looping Effects of Psychiatric Classifications: What Is an Interactive and Indifferent Kind?, *International Studies in the Philosophy of Science*, 21:329–344.

Tsou, J. Y. (2011), The Importance of History for Philosophy of Psychiatry: The Case of the DSM and Psychiatric Classification, *Journal of the Philosophy of History*, 5:446–470.

Tsou, J. Y. (2012), Intervention, Causal Reasoning, and the Neurobiology of Mental Disorders: Pharmacological Drugs as Experimental Instruments, *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43:542–551.

Tsou, J. Y. (2013), Depression and Suicide Are Natural Kinds: Implications for Physician-Assisted Suicide, *International Journal of Law and Psychiatry*, 36:461–470.

Tsou, J. Y. (2015), DSM-5 and Psychiatry's Second Revolution: Descriptive vs. Theoretical Approaches to Psychiatric Classification, pp. 43–62. In *The DSM-5 in Perspective: Philosophical Reflections on the Psychiatric Babel*. S. Demazeux and P. Singy, eds. Dordrecht: Springer.

Tsou, J. Y. (2016), Natural Kinds, Psychiatric Classification, and the History of the *DSM*, *History of Psychiatry*, 27:406–424.

Whooley, O. (2010), Diagnostic Ambivalence: Psychiatric Workarounds and the Diagnostic and Statistical Manual of Mental Disorders, *Sociology of Health & Illness*, 32:452–469.

Wilson, M. (1993), DSM-III and the Transformation of American Psychiatry: A History, *American Journal of Psychiatry*, 150:399–410.

Young, A. (1995), *The Harmony of Illusions: Inventing Post-Traumatic Stress Disorder*, Princeton, NJ: Princeton University Press.

Zachar, P. (2000), Psychiatric Disorders Are Not Natural Kinds, *Philosophy, Psychiatry, & Psychology*, 7:167–182.

10

Inductive Risks and Psychiatric Classification

Aaron Kostko

Debates about the role of values in psychiatry have tended to focus almost exclusively on their role in definitions of mental disorder and psychiatric diagnosis (Boorse 1976; Fulford 1989; Sadler 2005; Wakefield 1992). Recently, however, philosophers of psychiatry have started to focus more explicitly on their role in psychiatric explanation (Campaner 2014; Kendler 2014; Van Bouwel 2014) and the intersection between psychiatric research and classification (Cooper 2016; Frances 2013; Frances and Jones 2012; Ghaemi 2010, 2013; Wakefield 2010). This shift in focus has highlighted questions about how to balance the various epistemic goals of psychiatric research, such as constructing a valid classification of psychiatric conditions, and the nonepistemic goals toward which much of psychiatric research aims, such as alleviating harms to patients. Moreover, it has raised questions about the extent to which psychiatric researchers have a moral responsibility to consider the nonepistemic consequences associated with their research.

This chapter examines two broad approaches that one can take with respect to these two issues. The first approach, what I refer to as the value-free ideal of psychiatry, advocates for insulating psychiatric research and classification from the influence of nonepistemic values and a minimalist view of the moral responsibilities of psychiatric researchers. The second approach, what I refer to as the value-laden ideal of psychiatry, advocates for the inclusion of nonepistemic values in psychiatric research and classification decisions and a more expansive view of the moral responsibilities of psychiatric researchers. I outline the basic arguments in favor of each approach and assess their relative strengths and weaknesses. Although each approach has obstacles that they must overcome, I argue that it is not possible to completely insulate psychiatric research and classification from the influence of nonepistemic values and, therefore, that psychiatric researchers are responsible for considering the

nonepistemic consequences associated with their work. I begin by summarizing the distinction between epistemic and nonepistemic values and how they may conflict in the context of psychiatric research and when making decisions about how to revise diagnostic categories. I then outline and evaluate the arguments for the value-free and value-laden ideals of psychiatry, respectively.

Conflicts between epistemic and nonepistemic values

The distinction between epistemic and nonepistemic values provides a natural starting point for thinking about how to balance the various epistemic and nonepistemic goals of psychiatric research. Daniel Steel (2010) provides a useful framework for thinking about the differences between these two types of values. On his account, nonepistemic values are those values that aim at the attainment of some particular practical, moral, or social good. Common examples of these goods include the minimization of harms, justice and equity, respect for autonomy, social utility, economic goods such as job creation or profit, and personal goods such as prestige or recognition. There is broad consensus that values such as the minimization of harms, justice and equity, respect for autonomy, and social utility may play a legitimate role when pursuing the nonepistemic goals of psychiatry, such as deciding which psychiatric conditions warrant investigation, which research programs to fund, which research designs are ethically permissible, and how to apply research results to develop more efficacious treatments. These goals are typically considered to be "external" to psychiatric research. However, as will be discussed below, there is less agreement as to whether and to what extent these sorts of values should play a role in what one might refer to as the "internal" aspects of psychiatric research, such as decisions about how to interpret research results or whether to accept or reject a hypothesis.

Epistemic values, on the other hand, are those values that aim to promote the attainment of truth. Some values, which Steel refers to as intrinsic epistemic values, are necessary for and directly promote the attainment of truth. Examples of such values include empirical adequacy, predictive accuracy, and internal consistency. The idea is that an explanation that is not empirically adequate, fails to make accurate predictions, or contradicts itself is unlikely to be true. To use an example more directly related to the present discussion, a classification system that does not accurately characterize the symptoms of a condition, fails to make generalizable predictions regarding diagnoses, or leads to contradictory

diagnoses is unlikely to be accepted as valid. Other values, which Steel refers to extrinsic epistemic values, indirectly promote the attainment of truth by supporting or enabling the realization of intrinsic epistemic values. Examples of extrinsic epistemic values include testability, simplicity, generality, external consistency, and perhaps fundability. The idea is that an explanation or classification that fails to embody at least some of these values will likely fail to realize intrinsic epistemic values and, therefore, fail to promote the attainment of truth. Although there is disagreement about which values should count as epistemic values, how best to interpret each value, and the relative significance to place on each value, there is broad consensus that one should generally try to maximally realize epistemic values when pursuing the epistemic goals of psychiatry, such as constructing a valid classification of psychiatric conditions or identifying the biological mechanisms underlying psychiatric conditions.

Despite this consensus, there is less agreement as to whether nonepistemic values should influence the pursuit of these epistemic goals. For instance, to what extent should psychiatric researchers consider nonepistemic consequences when proposing revisions to diagnostic categories or criteria? Since psychiatric researchers will often have insufficient evidence to determine conclusively whether to revise a diagnostic category or criterion, any proposed revision will inevitably involve the possibility of error. Philosophers of science refer to this possibility of error as inductive risk. Hempel (1965) first introduced the term to refer generally to the possibility of being wrong when accepting or rejecting scientific hypotheses. The term has since been invoked to illuminate the risks in a wide variety of research contexts, such as the regulation of dioxin levels (Cranor 1993, 2006; Douglas 2000, 2009), environmental justice and global climate change (Steel and Whyte 2012), paleoanthropology (Elliott and Willmes 2013), divorce (Andersen 2004), and disorders of consciousness (Johnson 2016).

In the context of psychiatry, the inductive risks associated with diagnostic revisions can have widespread nonepistemic consequences for a variety of stakeholders, most notably for patients. For instance, if one infers from one's research that it is best to lump or split a diagnostic category or to revise the diagnostic criteria for a specific category and one is wrong, then one runs the risk of increasing false positive or false negative diagnoses. False positive diagnoses occur when one diagnoses a patient as having a condition when the condition is not present. False negative diagnoses occur when one diagnoses a patient as not having a condition when in fact the condition is present. An increase in false positive diagnoses risks "medicalizing" normal problems of living, treating individuals with potentially harmful and expensive therapies, and stigmatizing

individuals who are not actually suffering from a disorder. These risks extend beyond potential harms to the individual and include potential harms to society in terms of increased health care costs. An increase in false negative diagnoses, on the other hand, risks prolonging individuals' suffering since they will not receive potentially beneficial treatments. As with over-diagnosis, the potential risks associated with under-diagnosis extend beyond harms to the individual. Diminished work productivity and, depending on the nature of the disorder, threats to public safety represent just some of the potential harms to society.

These sorts of concerns generated contested debates prior to the release of the *DSM-5*. There were disagreements about how to balance the relative risks of false positive and false negative diagnoses associated with proposed revisions to the diagnostic criteria for numerous psychiatric conditions, including body dysmorphic disorder (Phillips et al. 2010), the paraphilias (First and Frances 2008), bipolar disorder (Frances 2013; Frances and Jones 2012; Ghaemi 2010; Phelps and Ghaemi 2012), posttraumatic stress disorder (Forbes et al. 2011), major depressive disorder (Frances 2010), and autism spectrum disorder (Frances 2010), among numerous others. The most notable of these debates regarded the inclusion or exclusion criteria for bipolar disorder (Frances 2013; Frances and Jones 2012; Ghaemi 2010, 2013). Bipolar disorder, which became an official diagnostic category in the *DSM*-III (APA 1980), was split in the *DSM*-IV into bipolar I and bipolar II (APA 1994). The rationale for the split was to be able to more reliably distinguish genuine differences in patients based upon the severity of manic episodes. However, this decision had the effect of blurring the line between bipolar II, which is marked by less severe manic episodes than bipolar I, and major depressive disorder (Frances and Jones 2012). Several proposed revisions to the *DSM*-5 diagnostic criteria for bipolar I and II were intended to address this problem. One relatively uncontroversial revision required the presence of elevated levels of energy and activity in addition to elevated mood for both the manic episodes associated with bipolar I and the hypomanic episodes associated with bipolar II. The other two proposed revisions, however, were highly scrutinized. The first of these revisions allowed for the diagnosis of bipolar I and/or bipolar II if the manic and/or hypomanic episodes are induced by the use of antidepressants. The second revision, which was ultimately not accepted, proposed reducing the current required duration of a hypomanic episode for bipolar II from four days to two days.

There are clear nonepistemic consequences associated with each of these revisions. If the presence of an antidepressant-induced hypomanic episode is in fact a reliable diagnostic marker for bipolar II, then this revision decreases the

risk of false positive diagnoses of major depressive disorder, decreases the risk of false negative bipolar II diagnoses, and increases the likelihood that individuals will receive the most appropriate treatment to alleviate their suffering. If, on the other hand, the milder hypomanic episodes that accompany bipolar II cannot be reliably distinguished from the periodic antidepressant-induced elevated moods associated with major depressive disorder, then bipolar II will be over-diagnosed, major depressive disorder will be under-diagnosed, and individuals will either not receive treatment or be prescribed treatment that further exacerbates their symptoms. Similarly, if the appropriate duration of a hypomanic episode for bipolar II is two days instead of four, then individuals with bipolar II disorder will not be mistakenly diagnosed with major depressive disorder and will receive the most appropriate treatment to alleviate their suffering. If this shorter duration is mistaken, however, then individuals with major depressive disorder will be mistakenly diagnosed with bipolar II and potentially receive treatment that could further exacerbates their symptoms.

Value-free psychiatry

If psychiatric researchers should consider these potential nonepistemic consequences when interpreting the results of their research, then they may be able to help mitigate any potential harms associated with over- or under-diagnosis. However, allowing nonepistemic values to play a role in these internal stages of psychiatric research raises several concerns, many of which have been raised with regard to the role of nonepistemic values in scientific research more generally (Anderson 2004; Betz 2013; Brown 2013; Douglas 2000; Mitchell 2004). The most notable of these concerns is the problem of wishful thinking: if psychiatric researchers are expected to appeal to nonepistemic values when interpreting the results of their research, then any revisions to diagnostic categories may be more a reflection of researchers' value judgments than the actual underlying structure of psychiatric conditions. This would threaten the reliability and integrity of psychiatric classification and diagnosis since it would allow one to evaluate evidence or accept or reject hypotheses based upon social, moral, or political values rather than epistemic values. In short, there would be nothing to restrict researchers from utilizing evidence or accepting hypotheses simply because they supported their desired conclusion.

Ghaemi (2010, 2013) expresses this worry in his defense of both of the proposed revisions to the diagnostic criteria for bipolar II disorder discussed

above. He acknowledges that psychiatric researchers may legitimately consider nonepistemic consequences "as a lower-level priority" when decisions about how to revise diagnostic categories or criteria are undetermined by the evidence (2013: 51). However, Ghaemi contends that researchers would be mistaken to consider the potential nonepistemic consequences that could result from their research when such decisions can be decided solely by appeal to epistemic values, which he argues is the case with respect to both of the proposed revisions to the diagnostic criteria for bipolar II disorder. Although each of these revisions would likely increase bipolar II diagnoses and could potentially increase false positive diagnoses, Ghaemi argues that they are supported by the evidence and should not be discarded simply because of concerns about possible false positive diagnoses. To do so, he claims, "would only ensure that our diagnostic system will continue to reflect the opinions of its writers and be a consciously artificial system" (2010). Moreover, if concerns about false positive diagnoses are going to influence the interpretation of research results, then Ghaemi suggests that we should be equally concerned about false positive diagnoses of major depressive disorder, which he suggests would likely result if the *DSM*-5 retained the longer duration period.

The second worry with allowing nonepistemic values to play a role in the internal stages of psychiatric research is that it would be inconsistent with the principles of a liberal democratic society (Betz 2013). It might potentially enable researchers to unjustly impose their view about what constitutes acceptable levels of risk on the rest of society. At the very least, it would allow researchers' values to be disproportionately represented in any resulting classification or public policy decisions, which would potentially violate patients' autonomy and undermine democratic decision-making processes. This outcome could be particularly problematic if researchers do not have any special expertise in deciding what constitutes an acceptable level of risk.

Defenders of the value-free ideal try to avoid these concerns by advocating for the exclusion of nonepistemic values from the internal stages of psychiatric research altogether. If one can decide how to interpret data, whether to accept or reject hypotheses, or whether to revise diagnostic categories or criteria based solely upon epistemic values, then the problem of wishful thinking vanishes. If one cannot make such decisions based solely upon epistemic values, then defenders of the value-free ideal contend that one should simply remain agnostic on the matter and wait for further evidence to accumulate. In either case, the risks of error would be negligible or nonexistent, and researchers would have no need to consider the nonepistemic consequences of their work.

Value-laden psychiatry

Critics of this value-free conception of psychiatry acknowledge the force of these concerns, but argue that excluding nonepistemic values is either unattainable or undesirable. Those who adopt the former strategy either challenge the distinction between epistemic and nonepistemic values upon which the conception rests (Longino 1996; Rooney 1992) or contend that individuals are simply incapable of excluding nonepistemic values from their assessments of the adequacy of evidence (Biddle 2013; Colombo et al. 2016). Those who adopt the latter strategy argue that the value-free ideal is undesirable typically acknowledge the distinction between epistemic and nonepistemic values; however, they maintain that the variety of inductive risks involved throughout the various stages of research requires researchers to consider the nonepistemic consequences of their work (Biddle 2013; Brown 2013; Douglas 2000, 2009; Elliott and Willmes 2013; Steel 2010; Steel and Whyte 2012). Even if it were possible to remain agnostic and wait for further evidence to accumulate on certain matters in research contexts, it is often not practical to do so in clinical and policy-making contexts.

Although defenders of the argument from inductive risk offer slightly different accounts of the circumstances under which nonepistemic values should exert their influence and the extent of this influence, they all typically appeal to some version of the underdetermination thesis to justify the inclusion of nonepistemic values. Pierre Duhem (1914) and W.V.O. Quine (1951) first articulated the underdetermination thesis with respect to scientific theories. They argued that scientific hypotheses could never be tested in isolation because auxiliary assumptions, for example assumptions about the way the world works, which instruments to use to test hypotheses, whether the instruments are functioning properly, are always needed for theories to generate testable hypotheses. Given this holistic nature of hypothesis testing, when a hypothesis seems to be falsified by the data, it is not possible to determine whether this is due to the hypothesis itself or to one of the auxiliary assumptions. Moreover, it is always possible to modify a theory so that is consistent with the data or to take two logically incompatible theories and modify their auxiliary hypotheses so that they are equally consistent with the data. For these reasons, Duhem and Quine argued that theory choice was underdetermined. Contemporary appeals to underdetermination are less reliant on Duhem and Quine's view about the holistic nature of hypothesis testing and instead focus on the fact that researchers rarely have sufficient evidence to determine conclusively whether to accept or reject a hypothesis. Given this gap between evidence and hypothesis, any decision

to accept or reject a hypothesis will inevitably involve inductive risk due to the possibility of error. Researchers will not be able to rely solely upon epistemic values to minimize the risks associated with these possible errors. When this is the case, and especially when the risks are significant, defenders of the argument from inductive risk contend that nonepistemic values should influence the standards of evidence required for accepting or rejecting hypotheses.

This strategy seems to underlie the *DSM*-5 guidelines' request for researchers to address the relative harms associated with any proposed revisions or new diagnostic categories (Kendler et al. 2009). Frances (2012, 2013) also adopts this strategy in response to Ghaemi's arguments in favor of both of the proposed revisions to the diagnostic criteria for bipolar II. He questions whether clinicians are even capable of clearly distinguishing the milder hypomanic episodes that accompany bipolar II from the periodic elevated moods that can accompany major depressive disorder, and argues that this ability is even further diminished by the fact that antidepressants often induce periodic elevated moods in depressed patients. He is also skeptical of clinicians' ability to clearly distinguish between the antidepressant-induced hypomanic episodes and the return of normal mood in patients with major depressive disorder. If these diagnostic decisions are underdetermined by the presentation of symptoms in clinical contexts, then there is little reason to suspect that they could be avoided when researchers have to draw similar inferences in research contexts. Frances suggests that this sort of underdetermination is ubiquitous in psychiatric diagnosis and that setting any diagnostic threshold is inevitably arbitrary, a feature that is particularly problematic given the general tendency of researchers to be biased toward making recommendations that are more tolerant of false positive diagnoses. With regard to the proposed revisions to bipolar II, Frances worries that this bias would lead to more individuals with depression being diagnosed with bipolar II and being treated with antipsychotics and mood stabilizers, which would unnecessarily expose them to potential risks such as weight gain and diabetes (2013). These sorts of nonepistemic consequences lead Frances to advocate for a more conservative approach regarding diagnostic revisions.

Cooper (2016) shares Frances's worries about the potential harms associated with the proposed diagnostic revisions to bipolar II, although her argument for the inclusion of nonepistemic values does not rely on underdetermination. Wishing to avoid the empirical debates regarding these revisions, she asks the reader to consider "a disorder where *empirical data regarding validators suggests that diagnostic criteria should be relaxed* … but where there is also robust evidence that the disorder is currently being overdiagnosed and medicated with

drugs with dangerous side effects" (109; my italics added). The contrast between the argumentative strategies of Cooper and Frances, along with other critics of the value-free ideal more generally, is important here. Whereas Frances claims that the *DSM* Task Force should consider the nonepistemic consequences of the revisions *because* the evidence is inconclusive, Cooper is arguing that they should consider the nonepistemic consequences *even if* the evidence "suggests that diagnostic criteria should be relaxed." This difference has important implications for when one can legitimately appeal to nonepistemic values. On Frances's view, one can appeal to nonepistemic values only when the evidence is inconclusive and one's decision has nonepistemic consequences. However, on Cooper's view, one can legitimately appeal to nonepistemic values even when the evidence is conclusive.

This difference can have important implications for how well one can respond to the problem of wishful thinking. Proponents of the argument from inductive risk typically try to avoid the problem of wishful thinking by restricting the contexts in which nonepistemic values can legitimately function. Douglas (2000) and Steel (2010), for instance, argue that nonepistemic values should not take precedence over epistemic values or directly determine whether scientists accept or reject a hypothesis, as this would seem to sanction the kinds of wishful thinking that worries defenders of the value-free ideal. Rather, nonepistemic values should only function to determine how much evidence is required before accepting or rejecting a hypothesis or the appropriate level of statistical significance. For instance, if accepting a particular hypothesis could lead to significant harms to human health, then one should adopt a stricter standard of evidence before accepting the hypothesis. Conversely, if the risks of accepting a hypothesis are negligible, then one could justify adopting a more relaxed standard of evidence. Since nonepistemic values only play an indirect role in the internal stages of psychiatric research, they do not override epistemic values and, therefore, rule out the possibility of wishful thinking on the part of researchers.[1]

This strategy succeeds, however, only to the extent that decisions about whether to accept or reject a hypothesis are not fully determined by epistemic values. If one denies this, as Cooper seems to do in her hypothetical example, then it's not clear what prevents nonepistemic values from playing a direct role in the internal stages of psychiatric research, which makes her position more susceptible to the problem of wishful thinking. Cooper thinks this is not a significant concern for her view and offers two, mutually supportive arguments to illustrate why. She first appeals to Douglas's (2003, 2009) account of the

moral responsibilities of scientists to make the case for the undesirability of excluding nonepistemic values from psychiatric research. She acknowledges that constructing a valid classification of psychiatric conditions is a worthwhile epistemic goal that is constitutive of psychiatric researchers' and *DSM* task force members' role responsibilities. However, following Douglas, Cooper argues that this role responsibility does not override their general moral responsibilities of trying to foresee and avoid the potential harms associated with diagnostic revisions. The only way to avoid fulfilling these responsibilities is if there were adequate social structures in place to monitor and evaluate any potential harms. Cooper suggests that this not the case because psychiatric researchers and members of *DSM* task forces will typically be the only individuals who are sufficiently qualified to evaluate potential harms: "The reason why experts on psychiatric classification have a moral duty to consider harms that might result from changes to classification is because they are the only ones who possess the expertise required to judge some of those harms; it takes such experts to know whether some disorder is being overdiagnosed, for example" (106).

Even if one accepts that researchers must invoke nonepistemic values to fulfill their moral responsibilities, one still needs an account of how to prevent these values from leading to wishful thinking. Cooper's second argument is intended to address this concern. She contends that wishful thinking is unlikely to occur since there is widespread agreement about what constitutes a harm or a benefit: "At least in the current context, there is no realistic risk that starting to consider values when revising the *DSM* would open the door to those who would seek to shape the manual to fit unacceptable value systems; all those involved in the debates broadly agree on what counts as a 'harm' or 'benefit'" (109). Given this broad consensus, any values that would lead to the problem of wishful thinking will likely be screened out.

It is unlikely, however, that this response will satisfy defenders of the value-free ideal. Although Cooper is likely correct that there is broad consensus regarding the benefits of accurate diagnoses and harms of false positive or false negative diagnoses, this does not eliminate the possibility of other sorts of nonepistemic values, such as considerations of profit or prestige, from influencing decisions about how to revise diagnostic categories or criteria. Moynihan and colleagues. (2013) provide evidence that suggests this could be a genuine possibility. For instance, in their systematic review, they found evidence of a publication bias toward the expansion of disease categories as well as frequent financial ties between the authors of these publications and pharmaceutical companies. This does not establish a direct relationship between nonepistemic values such

as profit or prestige and proposed diagnostic revisions, but it is enough to be skeptical of Cooper's claim that "there is no realistic risk" that such direct relationships could occur.

The upshot of these considerations is that if one wants to avoid the problem of wishful thinking then one needs to appeal to some version of underdetermination and provide an account of how to restrict the contexts in which nonepistemic values can legitimately function. Douglas's distinction between the direct and indirect roles suggests at least a possible way to do this. However, even if nonepistemic values only play an indirect role in psychiatric research, the concern about researchers unjustly imposing their view about what constitutes acceptable levels of risk on the rest of society still remains. Critics of the value-free ideal can respond to this problem by either challenging the account of researchers' moral responsibilities that defenders of the value-free ideal implicitly or explicitly endorse or by suggesting alternative mechanisms that can minimize the possibility of researchers imposing their values on the rest of society.

Adherents of the value-free ideal typically defend a minimalist account of researchers' moral responsibilities, maintaining that it is not the responsibility of researchers to consider nonepistemic consequences in the context of their research. Mitchell (2004), for instance, adopts this strategy in defense of the value-free ideal of science. Contrary to Douglas, she insists on a sharp division between scientific research and public policy contexts. The former context aims to establish reliable beliefs about the world, such as what is the most valid classification of psychiatric conditions, whereas the latter aims to determine which actions to take or policies to adopt based upon these beliefs, such as how best to revise existing psychiatric classifications. Although these two contexts often overlap, Mitchell claims that they embody different standards for acceptance and are guided by different sorts of values, with research contexts adopting higher standards for acceptance and appealing only to epistemic values to determine which beliefs are reasonable to accept.

With respect to debates about proposed revisions to diagnostic categories, Mitchell's position would entail that psychiatric researchers should only consider epistemic factors when drawing inferences about diagnostic revisions, such as whether the proposed revision is more accurate, general, or consistent with other background knowledge than the alternative. Ghaemi endorsed a similar view in his defense of the proposed revisions to the diagnostic criteria for bipolar II disorder. It is only when one attempts to use the results accepted in research contexts to inform decisions about how revise existing diagnostic categories that

one should consider the nonepistemic consequences of such decisions. At this point, however, such decisions are the purview of *DSM* task force members and not the psychiatric researchers. Of course, psychiatric researchers also often serve as members on *DSM* task forces. When this is the case, such individuals will inevitably rely on both epistemic and nonepistemic values to inform decisions. However, Mitchell is clear that "this is because two sets of role obligations are involved and not because the two types of values are warranted in each of the individual roles of scientist and policy advisor" (253).

To the extent that Mitchell's strategy succeeds, it would only exempt psychiatric researchers from the responsibility to consider the nonepistemic consequences of diagnostic revisions. One would need further arguments to exempt *DSM* task force members from having a similar responsibility. Nonetheless, this would pose a problem for Cooper's account of researchers' moral responsibilities. Although she frequently oscillates between attributing the responsibility to consider the nonepistemic consequences of diagnostic revisions to psychiatric researchers and *DSM* task force members, she clearly states that the former are at least partially responsible for this task, which is what Douglas's original account was intended to establish. Thus, if successful, Mitchell's account would at least undermine Cooper's claim that psychiatric researchers are responsible for considering the nonepistemic consequences of diagnostic revisions.

Another possible, and perhaps complementary, strategy to insulate psychiatric research from nonepistemic values and to minimize researchers' moral responsibilities is to try to reduce the uncertainty associated with the sorts of inferences that psychiatric researchers draw. The force of the arguments from Frances and Cooper, and the argument from inductive risk more generally, hinges upon the assumption that the sorts of inferences that psychiatric researchers draw, such as whether the evidence supports revising a diagnostic category or criteria, involve significant uncertainties and risks. Only if this assumption is correct can nonepistemic values legitimately influence the standards of evidence required for accepting or rejecting hypotheses. If one can eliminate uncertainty, then there is no risk associated with decisions about whether to accept or reject hypotheses, and thus no need for researchers to invoke nonepistemic values in the context of their work.

Betz (2013) adopts this strategy in his defense of the value-free ideal of science. His account rests upon a distinction between plain and hedged hypotheses. Plain hypotheses are empirical statements that do not make explicit reference to uncertainties in their formulation and are typically accepted beyond a reasonable doubt. The examples that Betz gives of plain hypotheses, all of

which are drawn from research on global climate change, include statements such as the atmosphere comprises oxygen, coal burns, and CO_2 is a greenhouse gas. Although each of these hypotheses could turn out to be wrong, they have been well tested, involve negligible risks, and are accepted to be reliable in public policy decisions. Most importantly, Betz claims that their acceptance is based solely upon their epistemic merits and does not involve any appeal to nonepistemic values. Analogous examples of plain hypotheses relating to bipolar disorder might be seemingly trivial statements, such as "A diagnosis of bipolar II requires an individual to have at least one hypomanic episode and one depressive episode," "A diagnosis of bipolar II requires an individual to have a hypomanic episode that lasts for at least four days," or "The standard treatment for bipolar II is mood stabilizers and antipsychotic medications."

Hedged hypotheses, by contrast, make explicit reference to uncertainties in their formulation. Betz mentions three ways that hedged hypotheses make uncertainties explicit. First, hedged hypotheses can report ranges of probability, for example "Hypomanic episodes last only 2–3 days in 27–30 percent of individuals." Second, they can be formulated as epistemically qualified statements, for example "It is likely/probable/plausible/consistent with what we know that hypomanic episodes last only 2–3 days in 27–30 percent of individuals." Finally, they can be formulated as conditional statements, for example "If we accept that hypomanic episodes last only 2–3 days in 27–30 percent of individuals, then requiring a shorter duration period for hypomanic episodes is preferable given our concern for minimizing false negative diagnoses." Provided that hedged hypotheses state ranges of probability or are appropriately qualified or conditionalized, Betz contends that they are no different from plain hypotheses: although they could be wrong, they involve negligible risks and can be accepted beyond a reasonable doubt to guide public policy decisions. Most importantly, their acceptance need not invoke any nonepistemic values. If one is concerned that this is not the case, then Betz suggests that one can further hedge one's hypothesis until it does possess the same error probabilities as other commonly accepted plain hypotheses.

Betz's strategy has important implications not only for the fate of the value-free ideal of science, but also for the responsibilities of psychiatric researchers. Since the acceptance or rejection of hypotheses need not invoke nonepistemic values, researchers only have the responsibility to report any error probabilities associated with the hypotheses that they accept or reject.[2] They would not be responsible for considering the potential impact of these error probabilities when used outside of research contexts, such as the impact that they may

have on false positive or false negative diagnoses. This account of the moral responsibilities of researchers stands in stark contrast to the views of Douglas and Cooper outlined above. Recall Cooper's remark that psychiatric researchers and *DSM* task force members "are the only ones who possess the expertise required to judge some of those harms; it takes such experts to know whether some disorder is being over diagnosed, for example" (106). While psychiatric researchers and *DSM* task force members are likely the most qualified to judge whether some disorder is being over-diagnosed, it is less clear whether they are the most qualified to judge the harms associated with this over-diagnosis. For instance, one might question whether psychiatric researchers and *DSM* task force members have any particular expertise to judge over-diagnosing bipolar II as more harmful than potentially underdiagnosing major depressive disorder. If Betz's strategy for hedging hypotheses is successful, then it eliminates this possibility and prevents researchers from imposing their standards of acceptable levels of risk on the general public. Of course, it remains an open question as to whether *DSM* task forces or other external agencies that are responsible for considering the nonepistemic consequences of diagnostic revisions are sufficiently representative and/or democratically authorized, but Betz's strategy would at least absolve psychiatric researchers of such a responsibility.

If the strategies of Mitchell and Betz succeed, then there would seem to be no justification for nonepistemic values to play a role in psychiatric research or for holding psychiatric researchers responsible for considering the nonepistemic consequences of their work. However, there are several reasons to be skeptical about whether their proposals can fully insulate psychiatric research and classification from nonepistemic values. Consider first Mitchell's proposal. Her position succeeds to the extent that psychiatric researchers will be able to determine the most valid classification of psychiatric conditions solely by appeal to epistemic values. Mitchell, however, does not provide an account of how this occurs. She acknowledges that scientists can disagree about how to interpret and/or apply various epistemic values, but she nonetheless insists that one determines which beliefs to accept as reasonable exclusively by appeal to epistemic values. For instance, she states: "While evidence underdetermines the truth of a causal claim, the other epistemic values … or cognitive values … can be harnessed to generate a judgment of acceptance" (249). Later she makes a similar point: "Scientists can certainly disagree … about the significance of different epistemic values or the degree of their satisfaction—without nonepistemic values entering in" (253).

If scientists can disagree about the interpretation of or particular weights attached to different epistemic values, then Mitchell owes an account of why

different researchers choose the weighted set of interpreted values that they do without appealing to nonepistemic values. For instance, two groups of researchers may accept as accurate research indicating that antidepressant-induced mania is 2.6 times higher in individuals diagnosed with bipolar disorder than in individuals diagnosed with major depressive disorder. However, they may disagree about whether this difference is sufficient to draw the inference that the presence of an antidepressant-induced hypomanic episode is a reliable diagnostic marker for bipolar II. Perhaps there is a deeper disagreement among the two groups about how to interpret accuracy or perhaps one of the groups thinks that the differential rate of antidepressant-induced mania among patients with bipolar II and major depressive disorder is not sufficiently generalizable to use it as a diagnostic marker. Those who wish to insulate psychiatric research from nonepistemic values need to provide an account of the source of the disagreement that avoids any reference to nonepistemic values. If one can only explain the source of the disagreement by appealing to nonepistemic values, then such values are still operative in the internal stages of research. So long as this is the case, then psychiatric researchers would still seem to be morally responsible for considering the nonepistemic consequences of their preferred interpretation and weighting of epistemic values.

Betz's proposal faces similar challenges. His proposal succeeds to the extent that researchers only need to invoke nonepistemic values at the final stage of deciding whether to accept or reject hypotheses for public policy decisions. While this is certainly a stage at which hypotheses are accepted or rejected, his argument ignores the variety of decisions that researchers make throughout the research process that have potential nonepistemic consequences. For instance, in the context of research on the duration of hypomanic episodes for bipolar II, researchers make decisions about what the inclusion and exclusion criteria should be for participation in a study, the duration of the clinical trial, if and when to terminate a clinical trial, if and for how long to follow up with research participants after a trial has ended, whether participants groups are sufficiently representative, how to handle cases of attrition, whether participants' reports about the length and severity of their episodes are reliable, among others. Of course, the potential consequences associated with each of these decisions will vary in their severity or likelihood. Nonetheless, for Betz's argument so succeed, he would have to show how one can hedge hypotheses at each of these various decision points. Perhaps this is possible, but the resulting hypothesis that one would accept or reject would be much more complicated than something as simple as "Hypomanic episodes last only 2–3 days in 27–30 percent of

individuals." Researchers would have to include error probabilities for each decision point to fully eliminate any need to appeal to nonepistemic values. It is doubtful that researchers would actually be able to objectively determine such error probabilities, let alone do so in such a way that would render the resulting hypothesis useful for practical decision-making.

So long as defenders of the value-free ideal are unable to provide an account of how epistemic values can solely determine which diagnostic revisions to make or how hedging hypotheses can eliminate the inductive risks associated with such revisions, then it is not clear how one can exclude nonepistemic values from the internal stages of psychiatric research. Moreover, it will be difficult to exempt psychiatric researchers from the responsibility to consider the nonepistemic consequences associated with their work. Even if they are not necessarily the most qualified to judge the relative harms associated with diagnostic revisions, they do not have to be the only individuals making such judgments. Other groups such as advisory boards, institutional review boards, public policymakers, and the general public may also share such responsibilities. By ensuring that these various stakeholders are actively involved throughout the research process, one can minimize the possibility of researchers imposing their values on the rest of society. One can further alleviate this concern by encouraging, or even requiring, scientists to be more transparent about which nonepistemic values they are using when evaluating the reasonableness of risks or the adequacy of evidence, as this will allow citizens to provide feedback on scientists' decisions and to ensure that their own values are reflected in public policy. The *DSM*-5's public comment period on proposed revisions is an initial step toward realizing this suggestion. Although these proposals require further elaboration, they at least suggest possible ways to ensure that researchers' values are not disproportionately represented in any resulting classification or public policy decisions.

Notes

1 See Elliott (2013) and Hicks (2014) for more detailed criticisms of Douglas's account of direct and indirect roles for values in science.

2 Researchers, of course, still have to fulfill other general role responsibilities, for example not engaging in deception or fraud. Moreover, as with Mitchell's position, researchers may occupy dual roles of researcher and policymaker. When this is the case, one would be justified in appealing to both epistemic and nonepistemic values.

References

American Psychiatric Association (1980), Task Force on DSM-III. In *Diagnostic and Statistical Manual of Mental Disorders: DSM-III*, Arlington VA: American Psychiatric Pub Inc.

American Psychiatric Association (1994), Task Force on DSM-IV. In *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*, Arlington VA: American Psychiatric Pub Inc.

Anderson, E. (2004), Uses of Value Judgments in Science: A General Argument, with Lessons from a Case Study of Feminist Research on Divorce, *Hypatia*, 19:1–24.

Betz, G. (2013), In Defence of the Value Free Ideal, *European Journal for Philosophy of Science*, 3:207–220.

Biddle, J. (2013), State of the Field: Transient Underdetermination and Values in Science, *Studies in History and Philosophy of Science Part A*, 44:124–133.

Boorse, C. (1976), What a Theory of Mental Health Should Be, *Journal for the Theory of Social Behaviour*, 6:61–84.

Brown, M. (2013), Values in Science beyond Underdetermination and Inductive Risk, *Philosophy of Science*, 80:829–839.

Campaner, R. (2014), Explanatory Pluralism in Psychiatry: What Are We Pluralists About, and Why?, pp. 87–103. In *New Directions in the Philosophy of Science*. M. C. Galavotti, D. Dieks, W. J. Gonzalez, S. Hartmann, T. Uebel, and M. Weber, eds. Dordrecht: Springer.

Colombo, M., Bucher, L., and Inbar, Y. (2016), Explanatory Judgment, Moral Offense and Value-Free Science, *Review of Philosophy and Psychology*, 7:743–763.

Cooper, R. (2016), First Do No Harm? What Role Should Considerations of Potential Harm Play in Revising the DSM?, *Philosophy, Psychiatry, & Psychology*, 23:103–113.

Cranor, C. (1993), *Regulating Toxic Substances. A Philosophy of Science and the Law*, New York: Oxford University Press.

Cranor, C. (2006), *Toxic Torts: Science, Law and the Possibility of Justice*, Cambridge: Cambridge University Press.

Douglas, H. E. (2000), Inductive Risk and Values in Science, *Philosophy of Science*, 67:559–579.

Douglas, H. E. (2003), The Moral Responsibilities of Scientists (Tensions between Autonomy and Responsibility), *American Philosophical Quarterly*, 40:59–68.

Douglas, H. E. (2009), *Science, Policy, and the Value-Free Ideal*, Pittsburgh, PA: University of Pittsburgh Press.

Duhem, P. (1914/1954), *The Aim and Structure of Physical Theory*, trans. from 2nd edition by P. W. Wiener, Princeton, NJ: Princeton University Press.

Elliott, K. (2013), Douglas on Values: From Indirect Roles to Multiple Goals, *Studies in History and Philosophy of Science Part A*, 44:375–383.

Elliott, K. and Willmes, D. (2013), Cognitive Attitudes and Values in Science, *Philosophy of Science*, 80:807–817.

First, M. and Frances, A. (2008), Issues for DSM-V: Unintended Consequences of Small Changes: The Case of Paraphilias, *The American Journal of Psychiatry*, 165:1240–1241.

Forbes, D., Fletcher, S., Lockwood, E., O'Donnell, M., Creamer, M., Bryant R. A., McFarlane, A., and Silove, D. (2011), Requiring Both Avoidance and Emotional Numbing in DSM-V PTSD: Will It Help?, *Journal of Affective Disorders*, 130:483–486.

Frances, A. (2010), Opening Pandora's Box: The 19 Worst Suggestions for DSM-5, *Psychiatric Times*, 27.

Frances, A. (2013), DSM in Philosophyland: Curiouser and Curiouser, pp. 95–103. In *Making the DSM-5*. J. Paris and J. Phillips, eds. New York: Springer.

Frances, A. and Jones, K. D. (2012), Bipolar Disorder Type II Revisited, *Bipolar Disorders*, 14:474–477.

Fulford, K. (1989), Moral Theory and Medical Practice, Cambridge: Cambridge University Press.

Ghaemi, N. (2010), Couch-Pragmatism, *Psychiatry Times*, 27:16–21.

Ghaemi, N. (2013), Taking Disease Seriously: Beyond "Pragmatic" Nosology, p. 42. In *Philosophical Issues in Psychiatry* II: *Nosology*. K. Kendler and J. Parnas, eds. Oxford: Oxford University Press.

Hempel, C. (1965), Science and Human Values, pp. 81–96. In *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*, New York: The Free Press.

Hicks, D. (2014), A New Direction for Science and Values, *Synthese*, 191:3271–3295.

Johnson, L. (2016), Inference and Inductive Risk in Disorders of Consciousness, *AJOB Neuroscience*, 7:35–43.

Kendler, K. (2014), The Structure of Psychiatric Science, *American Journal of Psychiatry*, 171:931–938.

Kendler, K., Kupfer, D., Narrow, W., Phillips, K., and Fawcett, J. (2009), Guidelines for Making Changes to DSM-5. URL= www.dsm5.org/ProgressReports/Documents/Guidelines-for-Making-Changes-to-DSM_1.pdf.

Longino, H. (1996), *Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy*, pp. 39–58. In *Feminism, Science, and the Philosophy of Science*, L. H. Nelson and J. Nelson, eds. Dordrecht: Springer.

Mitchell S. D. (2004), The Prescribed and Proscribed Values in Science Policy, pp. 245–255. In *Science Values and Objectivity.* P. Machamer and G. Wolters, eds. Pittsburgh, PA: University of Pittsburgh Press.

Moynihan, R., Cooke, G. P. E., Doust, J. A., Bero, L., Hill, S., and Glasziou, P. P. (2013), Expanding Disease Definitions in Guidelines and Expert Panel Ties to Industry: A Cross-Sectional Study of Common Conditions in the United States, *PLoS Medicine*, 10:e1001500.

Phelps, J. and Ghaemi, N. (2012), The Mistaken Claim of Bipolar "Overdiagnosis": Solving the False Positives Problem for DSM-5/ICD-11', *Acta Psychiatrica Scandinavica*, 126:395–401.

Phillips, K. A., Wilhelm, S., Koran, L. M., Didie, E. R., Fallon, B. A., Feusner, J., and Stein, D. J. (2010), Body Dysmorphic Disorder: Some Key Issues for DSM-V, *Depression and Anxiety*, 27:573–591.

Quine, W. V. O. (1951), Two Dogmas of Empiricism, pp. 20–46. In *From a Logical Point of View*, 2nd edition. Cambridge, MA: Harvard University Press.

Rooney, P. (1992), On Values in Science: Is the Epistemic/Non-epistemic Distinction Useful?, *Philosophy of Science*, 1:13–22.

Sadler, J. (2005), *Values and Psychiatric Diagnosis*, vol. 2. Oxford: Oxford University Press.

Steel, D. (2010), Epistemic Values and the Argument from Inductive Risk, *Philosophy of Science*, 77:14–34.

Steel, D. and Whyte, K. P. (2012), Environmental Justice, Values, and Scientific Expertise, *Kennedy Institute of Ethics Journal*, 22:163–182.

Van Bouwel, J. (2014), Pluralists about Pluralism? Different Versions of Explanatory Pluralism in Psychiatry, pp. 105–119. In *New Directions in the Philosophy of Science*. M. C. Galavotti, D. Dieks, W. J. Gonzalez, S. Hartmann, T. Uebel, and M. Weber, eds. Dordrecht: Springer.

Wakefield, J. C. (1992), The Concept of Psychiatric Disorder: On the Boundary between Biological Facts and Social Values, *American Psychologist*, 47:373–388.

Wakefield, J. C. (2010), Misdiagnosing Normality: Psychiatry's Failure to Address the Problem of False Positive Diagnoses of Mental Disorder in a Changing Professional Environment, *Journal of Mental Health*, 19:337–351.

11

Causal Explanation in Psychiatry

Tuomas K. Pernu

Introduction

The central aim of our scientific endeavor is to give us an accurate picture of the causal structure of the world. Having a valid and precise understanding of real causal relationships lends them to manipulation and control, which is something that is useful across the disciplines. This aspiration is particularly strong in the health sciences, which aim to give us understanding of the causes and mechanisms of diseases to enable us to make efficient clinical and preventative interventions. Among the most important questions of philosophy of medicine should therefore be these: What is "causation," what is "causal explanation," and what is "causal efficacy"?

Although causal explanation occupies a central place in all the health sciences, psychiatry is special in that it is affected by the mind-body problem and the issue of mental causation, the question of how, or whether, the mental and the physical can interact with each other. This chapter provides tools with which to analyze these issues in psychiatry and discusses some ways of tackling these problems in the light of some recent developments in the philosophy of science. Although these problems are not easy, and they are particularly grave in the context of psychiatry, recent discussion has brought about some significant advances which have the potential to enhance our understanding of the scientific identity of psychiatry.

The mind-body problem materializes in psychiatric context as a tension between psychological and physiological ways of explaining mental disorders. Your stance on the issue of how the mental and the physical are related, and whether they are in a causal interaction, therefore bears directly on what sort of psychiatric research and what kinds of clinical interventions you are bound to favor. The following will outline the basic philosophical elements of this

problem. It is shown how having a solid, philosophically and scientifically backed up account of causation can help one to reconcile the intuition that mental states can function as causes and effects with the idea that reality is thoroughly physical. However, a problem for this view is also sketched. Holding on to the idea of genuine (autonomous) mental causation requires one to assume that mental states cannot be identified with neural (or bodily) states. Although such an assumption can find support from current psychiatry, there are also reasons for skepticism, as shown in the end.

Causation in science and medicine

What is scientific explanation?

Let us take a few steps back and look at the scientific enterprise from a distance. What is it that we are trying to do in science? What is the ultimate goal? There are many perfectly good and enlightening answers to this question, of course, but at the same time it is clear that one answer is particularly pertinent: we are trying to get an accurate picture of the causal structure of the world. Why? Because having the correct understanding of the actual causal relations obtaining in the world enables us to implement effective strategies to control and manipulate the world according to our needs and desires. If there is a pragmatic interest to science, as there surely is, then this must be it.

Naturally there is much more to science than searching for causal explanations. For example, it is not at all clear—or at least it would be highly controversial to claim—that such abstract fields as philosophy, logic, mathematics, or theoretical physics are in any way engaged in an enquiry aiming at giving us an understanding of the causal structure of the actual world. However, at the same time it is clear that a large part, if not all, of the sciences we deem "empirical" are doing exactly that, from chemistry to psychology, and from biology to economics. Moreover, medicine is a field of enquiry that fits this model particularly well, and in several, interconnected ways.

Causation is central to medicine: diagnosis, prevention, treatment

Etiology is a central notion in medicine. Etiology, as it is typically understood, is the study of the causes or origins of a disease, disorder, or a condition that is thought to require medical attention. If you give it a moment's thought, and decipher the semantics of the term in your head, you'll see that it bears two

different meanings or functions in medicine. First, etiology is an essential element of diagnostics: knowledge of where and how a particular condition has originated plays a crucial role in determining the identity of the condition. What's central to medicine, and what's of particular concern in psychiatry, is the demarcation between the pathological and the (physiologically/psychologically) "normal." Pathology, in turn, is the study of causes and origins of a disease or a disorder—of the external source (pathogens) of abnormal, dysfunctional state of an organ or organism. It should be clear, therefore, that having reliable information on the origins of the condition is diagnostically useful. Moreover, such information is useful beyond diagnosing particular cases, for secondly, and maybe more importantly, medicine is not concerned only with studying diseases, but also, and primarily, with preventing diseases and other conditions that we deem harmful. Naturally, the more accurate information we have on the causes of a particular condition, the more successful we'll be in preventing that condition from arising.

But there is another, perhaps even a more essential way in which causal information becomes entangled with medical practice. Let us use etymology as our guide again: the term "medicine" means the practice of providing attention and care to the sick and injured (Charen 1951). This is what medicine aims at, first and foremost: to alleviate suffering, both physical and mental, and, ultimately, to heal those who are suffering from diseases, disorders, injuries, or other conditions that are considered harmful. In other words, medicine aims at specific, rather tangible and concretely useful *effects*, both at personal and at populational levels. And this in turn makes the search for effective treatments and practices the main focus of medicine. There is thus a constant need to determine the effective interventions vis-à-vis a particular condition, and to set these carefully apart from those interventions that are not effective (or not *as* effective, vis-à-vis this particular condition). At one extreme this core nature of medicine materializes as a fight against quackery and "alternative medicine"—as a fight against ineffective or even harmful interventions, advertised as effective.

From this concern arises also the more general interest in the constant assessing of the efficacy of medical practices and in developing better tools for the accurate assessment of such practices—i.e., the very idea of "evidence-based medicine"—and, consequently, the ever-growing interest in the placebo effect, the need to separate the "really effective" medical interventions from the merely apparently effective ones. Medicine, it thus seems, is thoroughly engaged with the idea of causal efficacy.

Since the notion of causation is so central to medicine, among the most important questions of philosophy of medicine should be these: What is "causation," what is "causal explanation," and what is "causal efficacy"? These are important, and old, questions of metaphysics and philosophy of science, and one must be prepared to approach the depth and complexity of the problems associated with them with a humble attitude. At the same time, however, it would be wrong to let oneself fall into despair and conclude that such questions are too profound and difficult to merit any systematic attention. This would be not just unfortunate but ill-founded, for many recent developments in metaphysics and philosophy of science have in fact rendered these issues systematically tractable. Philosophy has progressed, and analyzing the issue of causal explanation in psychiatry is actually a particularly enlightening way to demonstrate this. Having a rigorous understanding of some of the basic ideas of recent philosophy of science and philosophy of mind is quite concretely useful in addressing many foundational questions of psychiatry.

What is causation?

Causation: Dependency vs production

Although there are a number of well-developed accounts of causation in current philosophy (agential, counterfactual, interventionist, physical process, probabilistic, regulative, etc.), one can make one very useful taxonomical distinction, a distinction that is particularly useful in the current context. On the one hand, there are accounts that stress the idea that causation is a matter of interdependence of events or variables. On the other hand, there are accounts that stress the idea that causation is a matter of physical process or production (Hall 2004). The first camp includes, most prominently, the counterfactual (Lewis 1973), interventionist (Woodward 2003), probabilistic (Eells 1991; Suppes 1970; Williamson 2004), and structural equation or causal modeling (Halpern 2016; Pearl 2000; Spirtes et al. 2000) accounts. The second camp includes the causal line (Russell 1948), conserved quantity (Dowe 1992, 2000; Salmon 1997), energy transference (Castañeda 1980; Fair 1979), mark transmission (Salmon 1984), physical force (Bigelow et al. 1988; Bigelow and Pargetter 1990), and trope persistence (Ehring 1997, 2003; Kistler 1998, 2006) accounts. There is a significant overlap between these views, of course, but this overlap is typically between accounts falling into the same category. Very rarely,

if ever, do insights from one account seep through the boundary into another account in the competing camp.

This distinction is particularly relevant in the context of psychiatry for the tension between psychological and physiological explanations can be seen to be rooted in this distinction, or something closely akin to it. Although there are various subtle differences in the theories, there is one central point of departure to draw attention to: the different attitudes the two separate types of theories of causation hold on the issue of causal locality (*cf.* Dowe 2004; Schaffer 2004). We will need to rely on a very crude and intuitive notion of locality here, but the basic idea should be relatively clear. What the dependency accounts require of a proper causal relation is merely that some parts of the world (or representations of them) are dependent on each other in a way specified by the theory (counterfactually, probabilistically, etc.). What the production accounts require, on the contrary, is that these parts of the world (and not their representations) are physically connected to each other (or that the causal dependency can always be traced back to such a physical, concrete connection). In other words, what the latter accounts require is that there is a continuous, unbroken chain of events from causes to effects; their differences lie in the ways they specify what this continuous chain of events really amounts to.

Which view of causation is more "scientific"?

The physical production view of causation could initially seem more solid and scientific, but such an appearance is deceiving. First, although it is intuitively credible to think that genuine causal relations are concrete, local interactions—this is something that the physicalistic, scientific worldview would seem to suggest—a broad range of intuitively clear cases of causal interaction do not trade on such ideas, at least not in any obvious way. For example, it seems perfectly right to say that the macro-economic decline in the 1930s, the Great Depression, caused a number of suicides across Europe and North America. It is not hard to think other examples of intuitively cogent causal claims that we find impossible to translate into the language of physical pushing and pulling. Moreover, omissions and absences can typically function as perfectly good causes and effects, but such things are "nowhere," *per definitionem*, and hence would seem to be ill-suited to figure in any local physical interactions. For example, it seems perfectly right to say that the patient's failure to take her medication caused her psychotic episode. If the failure, the absence of medication, is functioning as a cause in this scenario, it clearly cannot act as a local producer of the effect.

Secondly, and more importantly, the dependency view reflects better how causal reasoning occurs in a variety of scientific disciplines than the physical production view. Science works through, in most typical cases, by analyzing data sets to identify stable and recurring connections between different data points. What one tries to do, in other words, is to spot genuine causal relationships from empirical data by using various statistical methods. The pretheoretic causal framework applied in this sort of analysis relies on statistical (counterfactual, probabilistic) dependencies, not on concrete physical production. Looking at the actual scientific practice, it would therefore be more appropriate to hold the dependency view as the "more scientific" of the two.

However, there is a subtle connection between the two views that often fails to get as much attention as it deserves. The main philosophical problem here is that it seems that one needs to supplement the purely statistical analysis with other, more concrete information to reach reliable causal conclusions. Statistical analysis will of course deliver only statistical results, but causation is naturally something different to statistical correlation. In the health sciences in particular, simple statistical evidence is rarely taken to be enough to ground causal claims. To reach generally accepted causal conclusions, the statistical evidence needs to be supplemented with mechanistic understanding (*cf*. Russo and Williamson 2007). Smoking, for example, is generally thought to cause lung cancer not just because smoking and lung cancer are statistically linked, but because we know what the ingredients of cigarettes are, and we know how consuming them is mechanistically linked to changes at the cellular level that give rise to cancer. And now, when you start to analyze this reasoning, you easily fall back into viewing causation as some sort of local relationship of physical production. This tension, or an interplay, if you wish to see the connection in a more positive light, between these two views on causation can be thought to be particularly tangible in psychiatry where the issue of how mental disorders and their physiological underpinnings are connected is constantly and concretely present.

The interventionist account of causation

This caveat notwithstanding, let us now have a closer look at one well-defined account of causation. To make the discussion as precise as possible, the following will rely on an interventionist account of causation, a paradigmatic representative of the dependency view. There are many reasons to adopt this account. First, the view is precisely defined and widely studied in recent philosophy of science, and one could say that it has become the dominating view on causation in current

philosophy. Second, the account is closely connected to scientific practice, and it suits analyzing causal claims in the health sciences particularly well. Third, it has received attention in recent philosophy of psychiatry, and it is claimed that new solutions have been found for a number of fundamental problems in psychiatry with the help of this sort of understanding of causation (e.g., Campbell 2008a–b, 2009; Kendler 2011; Kendler and Campbell 2009; Woodward 2008).

The interventionist account of causation integrates many insights from different accounts of the dependency view, the agency, counterfactual, and structural equation accounts in particular. According to interventionism, causal claims are claims about results of hypothetical interventions in counterfactual scenarios (e.g., Halpern 2016; Pearl 2000; Woodward 2003). There are two basic elements to this account. First, you need to specify a set of variables (e.g., {*C, E*}) that constitutes the domain of entities or events under scrutiny. Second, you define a set of structural equations specifying the various dependencies of the variables in the domain (e.g., (if $C = 1$, then $E = 1$) and (if $C = 0$, then $E = 0$)). Causal relations are, therefore, simply patterns of correlations among the values of the variables under hypothetical changes in them. More precisely, a variable *C* is a cause of a variable *E* (in the given domain) just in case there is an intervention on the value of *C* that will result in a change in the value of *E*.

There are of course various technical details to this account, but this rough idea should be enough for outlining the basic issues we are faced with in psychiatry regarding causal explanation. However, two philosophical issues are worth mentioning. First, interventionism offers a nonreductionistic analysis of causation. Note that typical analyses are reductionistic in that they analyze the notion of causation fully in noncausal terms: for example the counterfactual account analyzes the causal relation to counterfactual dependencies, the probabilistic to probabilistic dependencies, the transference account to transference of energy, and so on. Intervention, however, is clearly a causal notion, and it is used to define what causation is. Secondly, interventionism is anthropocentric, or at least has a strong anthropocentric element to it: causal claims make sense only in contexts where the relevant interventions can be carried out. Some might object that this leaves a number of objective causal processes outside of the analysis—how should we account for the cosmic effects of black holes, for example?—and argue that we therefore should not accept interventionism as a complete, or final analysis of causation. Although it is important to be aware of both of these critical issues, it can be assumed that the analysis is still useful, in the current context at least, as the following will demonstrate.

Causation and "levels of reality"

Psychiatry and the mind-body problem

Understanding causal explanation is central to understanding how medicine works. In psychiatry, however, we are faced with the issue of causal explanation in a setting that is particularly challenging. Although all medicine is complex—in the sense of being multifactorial and multilayered—psychiatry occupies its own level of complexity: psychiatry is trying to understand the mental realm, and to do so by navigating in-between the mental and the physical views on reality. In other words, basically all of psychiatry is thoroughly entangled with the mind-body problem, and the issue of how (or whether) the mental and physical can interact.

Deciphering the relationship of the mental and the physical is notoriously difficult. However, this is yet another issue where actual philosophical progress has been made, and we are in a position to formulate the problem, or at least some relevant parts of it, more precisely than ever, and therefore also able to reach precise results with the potential for concrete applications. Having a proper understanding of these developments is crucial to understanding the foundations of psychiatry.

What is "mental" then, as opposed to "physical"? What is the distinction? Although there are a number of things to draw attention to (*cf*. Pernu 2017), the following two characteristics are particularly pertinent in this context: first, our mental states are meaningful, that is, they refer to things and events outside of themselves and outside of the brains that ground them; second, our mental states, sensations in particular, are accompanied by specific phenomenal, subjective content, a feeling of what it is to have that particular experience. In other words, intentionality and subjective consciousness are the central characteristics of our mental lives. Perhaps nothing is more real to us than the phenomenal and semantically meaningful content of our mental states. Yet, we find it difficult, if not utterly impossible, to explain them in terms we otherwise hold ultimately real—in terms of the objective and mechanistic sciences of physics and biology.

Nonreductive physicalism and the idea that the mental is multiply realized by the physical

Let us now simply take for granted that speaking in psychological terms is natural and useful for us, and that replacing this way of speaking with a thoroughly

physicalistic parlance is not feasible for us. But the question still remains: How exactly should we conceive of the mental as distinct from the physical? How should the difference between the two be understood for us to make sense of the connection that we also perceive to be there between the two? The typical way to think about the relationship in current philosophy is to conceive of the mental as being fully dependent on, yet distinct from the physical. How to have one's cake and eat it too? By holding that the mental is always physically realized—that whatever constitutes the physical basis of a particular mental state determines the occurrence of that mental state in its entirety—but also holding that the mental is multiply realized—that each mental state could have had an alternative physical basis. This is the basic idea of nonreductive physicalism, the dominant view in current philosophy of mind.

It is not difficult to appreciate this view, and it is easy to see how it can be at home in psychiatry. It's monistic: the world is ultimately physical, with no separate realms or entities. But it leaves room for the autonomy of psychology: the mental is dependent, but not identical or reducible to the physical. No wonder the view is popular; it seems to offer something for everyone. How, then, should the core idea, the thesis of multiple realizability, be understood? According to this view the same mental states—or in principle all higher-level, functional states—could, both in principle and in practice, appear in various different material constitutions. Different people share the same thoughts even if their brains are not identical, different species share similar mental functions even if they are biologically radically different, and computers and robots can behave intelligently, and they are designed to resemble us, yet their material constitution is completely different from ours. The idea of multiple realization seems very natural to us.

At the heart of this thesis is the idea that (psychological) functions are implementation independent. In computing, you can implement the same software in different hardware, as is often done, and in principle you could build powerful computers out of sticks and rocks, or so the thinking goes. It is indeed quite attractive to think of psychiatry as a science focusing on the software of the mind, and neurology on the hardware. It is in many ways an illuminating and apt analogy. Both ways of viewing the system (the mind-brain system) are right, there exist efficient interventions at both levels, and there is a neat division of labor where the usefulness of both types of engineer is recognized. You need a functional hardware to run the given software. But often the malfunctions that appear are malfunctions of the software, not of the hardware.

Let us now suppose that we can in this way account for the intuitive idea that there are various "levels of reality" or "levels or organization" out there. Molecules

are made of atoms, organelles out of molecules, cells out of organelles, organs out of cells, organisms out of organs, and so on. Minds, on this view, are simply another level of reality arising from the right sort of biological organization—brains and central nervous system. Note that this way of construing the levels of reality as a nested functional hierarchy incorporates the idea of multiple realizability: entities at one level retain their identity even under radical changes at the lower levels—like organisms retain their identities even though they go through constant changes at the cellular level. Thus, we can hold onto the idea that there are distinct, higher levels to reality, the mental realm among them, that are dependent on the lower levels that realize them, and are nothing "over and above" them.

Causation in nonreductive physicalism

If we accept this view, how, then, should we perceive causation? What sort of causal interactions does this view allow? The received view, and the main thrust of nonreductive physicalism, is the idea that the higher-level entities and events are genuinely causally efficacious. In fact, this is the reason for holding them "real": higher-level entities and events have irreducible causal powers, and therefore we are committed to granting them autonomous existence. At the same time, however, the grounding, physical level is thought to be causally complete. In other words, nonreductive physicalism holds that each physical level event that has a cause, has a sufficient, complete physical cause. In this way, we do not need any higher-level information to account for the events occurring at the fundamental physical level, but at the same time, the view maintains, no information confined to the fundamental physical level could be sufficient to account for the causal relationships obtaining at the higher levels.

Let M and M^* now be variables standing for higher, mental-level events, and let $N1$ and $N2$, and N^*1 and N^*2 be variables standing for their respective lower, neural-level realizers. Figure 11.1, the iconic diagram of nonreductive physicalism (*cf*. Fodor 1974), illustrates how this view perceives the relationships between these variables.

This view seems to fit well with actual psychiatric practice: no-one doubts that mental events, and mental disorders, are neurally grounded, and that the neurophysiological realm forms a complete system, but at the same time it is the psycho-social interactions that function as typical determinants of the phenomena that call for psychiatric attention.

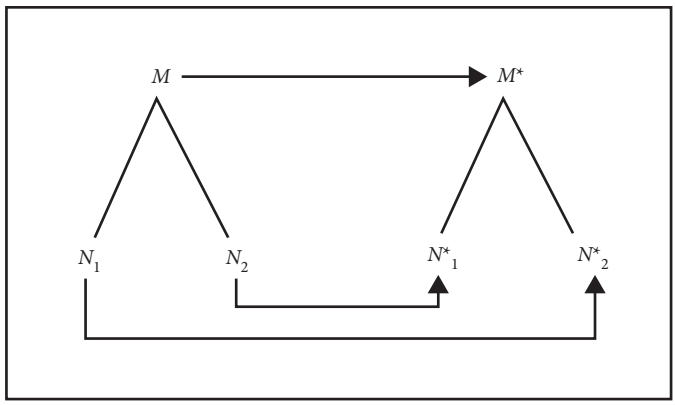


Figure 11.1 The iconic diagram of nonreductive physicalism. (Adapted from Fodor 1974).

It is wrong to pit social, psychological, and neural interventions against each other (even if one holds on to thoroughly physicalistic metaphysics)

Although the idea that mental states are neurally grounded should be uncontroversial, the direct consequences of this idea are easily underappreciated. First, it should be obvious that changes at the different levels of analysis, mental and neural, are correlated. Therefore, you cannot support claims on the primacy or autonomous existence of one level by relying solely on evidence showing that changes on this level result in changes on the other level (*cf*. Pernu 2011). For example, noting that certain psychological features have a specific cortical basis does not license the conclusion that those cortical features are causes of those psychological features (and hence that those psychological features are somehow less real). Or, conversely, the fact that psychological interventions (psychotherapy) result in changes at the neural level (cortical level), does not say anything about the efficacy of such interventions: we already knew that psychological interventions result in psychological changes, and we knew that mental states are neurally based, so we also know that the resulting psychological changes will always manifest as neural changes.

Another direct consequence of the sketched metaphysical view deserves to be highlighted: we can take it for granted that psychiatric disorders are typically multifactorial, which suggests that various different interventions can be effective, to varying degrees. It is therefore wrong to pit social, psychological,

and neural interventions against each other. For example, consider the fact that there is a strong positive correlation between experiencing abuse in childhood and suffering from various mental disorders (depression, anxiety) in adulthood. What sort of interventions should we apply to improve the situation? There is clearly no single right answer; it depends on what sort of effects you are aiming to produce, and at which point in the process you are able, or willing, to act. If you are a clinician, and you are faced with a patient with a unique history and personality, your interest is in finding an effective way of alleviating the current symptoms of this particular person. One effective way of doing that could be to apply neural-level interventions, that is, pharmacological treatment. But it would be wrong to conclude from this that mental disorders originating from childhood abuse are nothing but brain disorders and that pharmacological interventions are the only correct way to treat them. Obviously, if you are treating an adult patient with a condition that has resulted from childhood events, you are unable to intervene on the ultimate causes of the condition simply because you are unable to intervene on events in the past. However, if you are not faced with a particular patient, but the general issue of how to prevent these types of mental disorders from occurring, if, in other words, your effect-variable is a future-oriented population-level one, the recommended intervention should look totally different: you should intervene on the societal conditions that give rise to abusive behavior. All this should be rather obvious. Nevertheless, all too often psychiatric interventions targeting different levels of organization are treated as mutually exclusive. Neural-level interventions do not have to be the only appropriate interventions even if all mental states have, by necessity, a neural basis.

Are mental states really multiply realized?

Since the etiology of mental disorders is multifactorial, and the same types of disorders can rise through radically difference routes, it is natural to assume that mental disorders are multiply realized at the neural level. However, the idea of multiple realization cannot be taken for granted, especially in the context of current psychiatry.

There are a number of reasons to be critical of the thesis of multiple realization (e.g., Bechtel and Mundale 1999; Polger and Shapiro 2016; Shapiro 2000). The core of the criticism can be phrased in abstract metaphysical terms: if we deem

an entity multiply realized, then it would seem that its realizers would have to differ in their causal profiles; but if the realizers differ in their causal profiles, it is not clear why they should be treated as instances of the same multiply realized entity. Is it credible to assume that mental states can be just relevantly similar enough but also just relevantly different enough to be qualified as multiply realizable? A moment's reflection should make one suspect that such a position is hard to hold.

There is an internal tension in the multiple realizability thesis and that tension can be discharged in two different ways. On the one hand, the assumed multiply realized higher-level entity can split into separate entities each aligned with their realizers. On the other hand, the realizers that were assumed to be distinct can merge into a single realizer. Both of these possibilities dissolve the assumed multiple realization into identity (symmetrical dependence) between the higher- and lower-level entities (which would prompt one to be critical toward the whole stratified picture of reality). Figure 11.2 illustrates the situation.

If the problem would be merely conceptual or metaphysical in nature, the issue could be put aside and ignored. That is not the case, however. On the contrary, many of the fundamental issues we currently face in psychiatry are directly linked to the question of whether the thesis of multiple realizability applies to mental disorders. Moreover, this issue isn't merely theoretical. Different solutions to the issue have direct repercussions for how we position psychiatry among sciences, for what lines of research to undertake, and for what sort of clinical interventions to favor.

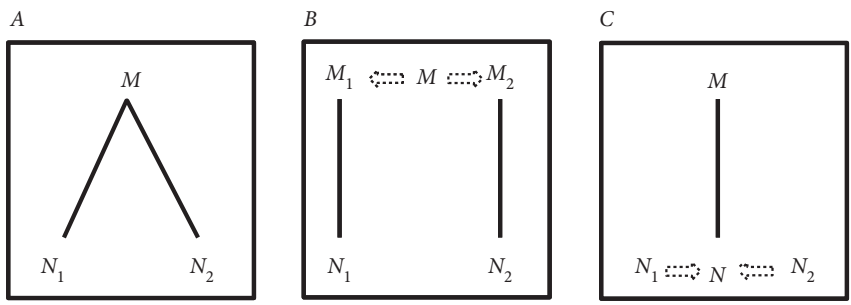


Figure 11.2 (A) represents the multiple realization hypothesis, (B) represent kind splitting, and (C) realizer unification or merging. (From Pernu (forthcoming).)

Two examples: Schizophrenia and addiction

Consider the following example. In autumn 2017 the British Psychological Society published an updated report on *Understanding Psychosis and Schizophrenia* (Cooke et al. 2017). This created a debate—not unexpectedly—on whether the report had taken all the relevant symptomatic, aetiological, and pharmacological issues into account. What this debate prompted is an editorial in *The Lancet Psychiatry* (Editorial 2018), urging us "to think of schizophrenia as an unmapped, ill-defined area, perhaps as an iceberg" (1). And as science progresses, the editorial predicted, we would witness how "minute but significant parts of this iceberg will break off" (Editorial 2018: 1). Now, whether or not you subscribe to this train of thought, you have to appreciate its foundations. The idea is that we don't know enough of the causes and effects of schizophrenia—or this thing we call "schizophrenia"—to properly understand its essence. And further, once we get there—once we have that understanding, or something close to it—we realize that there never really was such a thing, that there never was a unified, coherent mental disorder we should deem "schizophrenia." Our proper (concrete, physical, that is) understanding of the phenomenon has actually broken down and thus erased the phenomenon itself. There is no "schizophrenia," but only these separate chips of it, in the vein of Figure 11.2B.

Consider another, equally central issue in psychiatry: the problem of addiction. There is a debate about whether addiction should be construed as a choice or whether it should be construed as a (neurological) disease (e.g., Goldstein and Volkow 2011; Heyman 2009; Leshner 1997; Lewis 2015, 2017; Robinson and Berridge 2000; Szalavitz 2016; Volkow et al. 2016). But maybe this juxtaposition is ill-founded? A number of studies have indicated that only a minority of subjects develop a "pathological" substance addiction (e.g., Ahmed 2010; Deroche-Gamonet et al. 2004; Dutra et al. 2008; Robins et al. 1974). What this suggests, then, is that there might be different types of "addiction"—that addiction is not one coherent psychological or behavioral kind. These different types of addiction, in turn, would call for different types of interventions. So, to be more analytical, first, we observe that distinct clusters of subjects respond in distinct ways to addictive cues—distinct clusters of testees have distinct causal profiles vis-à-vis addictive scenarios. And second, we observe (or we should observe) that different types of interventions are efficacious, depending on which cluster the patient belongs to. This, again, should prompt us to split "addiction" into (at least two) different types, in accordance with their causal profiles. As a result, "addiction" would not be genuinely multiply realized, but would encompass

different things, each in alignment with their neural bases, in the vein of Figure 11.2B. Of course we might still have pragmatic reasons to keep using the one concept, and maybe even to treat the different types of addiction in the same institutions, but we should not let such social conventions mislead us into thinking that these different things are fundamentally the same.

The increasing pressure to dissolve the apparent multiple realization of mental disorders: The RDoC framework

To further motivate the thesis that psychiatry is in a particularly fragile state—"fragile" in the sense that its current nosological practices are under pressure to become more fine grained—consider the introduction of the Research Domain Criteria (RDoC) framework into psychiatric research (Cuthbert 2014; Cuthbert and Kozak 2013; First 2012; Insel et al. 2010). The traditional *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*) framework is based on symptomatic classification of disorders. This way of classifying mental disorders is prone to produce invalid results, the critics claim, for clusters of symptoms are not robust enough to provide us accurate and stable information about the disorders. That is why we should look into the various ways mental disorders are actually (biologically) realized and aim to classify them in a more accurate manner.

It is not difficult to appreciate the attractiveness of the RDoC framework. Although the framework is aimed primarily at improving research on mental disorders, it is clear that there is a heavy clinical thrust behind the initiative. In fact, one argument against the *DSM*, and in favor of the RDoC, has targeted the influence that the *DSM* has on drug development, and how that leads to suboptimal results due to the *DSM*-based categories lacking biological validity (e.g., First 2012; Hyman 2010). Having more accurate, valid classifications of mental disorders provides us with means to reach more accurate diagnosis, and, consequently, to apply more accurate and effective medical interventions. Having a better understanding of the biological constitution of a particular disorder makes us able to make better predictions of how the disorder behaves in different situations and under different interventions—in the same way that having correct understanding of the phylogenetic positioning of an organism allows us to make reliable predictions about the traits and behavior of that organism.

Whether this aim can be achieved depends on whether mental disorders are genuinely multiply realized—whether, at least in some important cases,

we are faced in reality with the Figure 11.2A rather than the Figure 11.2B (*cf*., e.g., Hoffman and Zachar 2017; Parnas 2014). If that is the case, then having appropriate information of the biological realizers of these mental disorders will not allow us to make optimally accurate and stable predictions. The jury is still out, and we will need more both conceptual and empirical research to reach the right verdict. However, it is clear that many indicators point to the conclusion that the idea of multiple realization of mental disorders cannot be taken for granted, and that many disorders now treated as homogeneous will split into separate disorders.

Concluding remarks

We are living exciting times in psychiatry, and especially in research on the foundational issues. There have been significant advances in the philosophy of science, especially on causation and on issues related to reduction and multi-level explanation. We can see more clearly than ever how our views on causation bear on how we think the mind and body are connected and this has potential to give us better understanding also on how the apparent tension between psychological and neural interventions could be discharged. At the same time, empirical research has developed more accurate tools, and more precise data has been accumulated on the biological underpinnings of mental disorders. This will lead, the hope is, to a more valid psychiatric nosology, with recognized disorders having more homogeneous causal profiles, which lends the disorders to more precise and efficient manipulation and control—to more precise and effective treatments. This hope might of course turn out to be ill-founded; we might have to admit that mental disorders are genuinely multiply realized at the physiological level, and that it is psychological, rather than pharmacological interventions, that are most effective in typical clinical cases. But whichever way we go the result will depend on our philosophical understanding on causation and multi-level explanation.

It is therefore not a stretch to claim that we are starting to have the elements in place which will enable us to gain significant advances in understanding the nature of mental disorders, if only the elements are put together the right way. To make that step, and to make that decisive advancement, what is needed is a more intimate collaboration between the conceptual and the empirical strains of research. Intellectual silos should therefore be abandoned, and more resources should be allocated to genuinely interdisciplinary research.

Acknowledgments

I would like to thank Dr. Robyn Bluhm, Mr. Peter Cave, Dr. Nadine Elzein, Dr. Outi Mantere, Dr. Hane Maung, Mr. Ivo Storrie, and Dr. Şerife Tekin, the participants of the Philosophy & Medicine Reading Group at King's College London in the winter/spring of 2017, and the participants of the Philosophy of Psychiatry Seminar at King's College London on January 22, 2018, for helpful discussions on the topic and comments on various versions of this chapter. This work has been financially supported by The Finnish Academy of Science and Letters, the Emil Aaltonen Foundation, and the Waldemar von Frenckell foundation.

References

Ahmed, S. H. (2010), Validation Crisis in Animal Models of Drug Addiction: Beyond Non-Disordered Drug Use toward Drug Addiction, *Neuroscience & Biobehavioral Reviews*, 35:172–184.

Bechtel, W. and Mundale, J. (1999), Multiple Realizability Revisited: Linking Cognitive and Neural States, *Philosophy of Science*, 66:175–207.

Bigelow, J., Ellis, B., and Pargetter, R. (1988), Forces, *Philosophy of Science*, 55:614–630.

Bigelow, J. and Pargetter, R. (1990), *Science and Necessity*, Cambridge: Cambridge University Press.

Campbell, J. (2008a), Causation in Psychiatry. In *Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology*. K. S. Kendler and J. Parnas, eds. Baltimore, MD: Johns Hopkins University Press.

Campbell, J. (2008b), Comment: Psychological Causation without Physical Causation. In *Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology*. K. S. Kendler and J. Parnas, eds. Baltimore, MD: Johns Hopkins University Press.

Campbell, J. (2009), What Does Rationality Have to Do with Psychological Causation? Propositional Attitudes as Mechanisms and as Control Variables. In *Psychiatry as Cognitive Neuroscience: Philosophical perspective*s. M. Broome and Lisa Bortolotti, eds. Oxford: Oxford University Press.

Castañeda, H. (1980), Causes, Energy and Constant Conjunctions. In *Time and Cause*. P. Van Inwagen, ed. Dordrecht: Reidel.

Charen, T. (1951), The Etymology of Medicine, *Bulletin of the Medical Library Association*, 39:216–221.

Cuthbert, B. N. (2014), The RDoC Framework: Facilitating Transition from ICD/DSM to Dimensional Approaches That Integrate Neuroscience and Psychopathology, *World Psychiatry*, 13:28–35.

Cuthbert, B. N. and Kozak, M. J. (2013), Constructing Constructs for Psychopathology: The NIMH Research Domain Criteria, *Journal of Abnormal Psychology*, 122:928–937.

Deroche-Gamonet, V., Belin, D., and Piazza, P. V. (2004), Evidence for Addiction-Like Behavior in the Rat, *Science*, 305:1014–1017.

Dowe, P. (1992), Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory, *Philosophy of Science*, 59:195–216.

Dowe, P. (2000), *Physical Causation*, New York: Cambridge University Press.

Dowe, P. (2004), Why Preventers and Omissions Are Not Causes. In *Contemporary Debates in Philosophy of Science*. C. Hitchcock, ed. Oxford: Blackwell.

Dutra, L., Stathopoulou, G., Basden, S. L., Leyro, T. M., Powers, M. B., and Otto, M. W. (2008), A Meta-Analytic Review of Psychosocial Interventions for Substance Use Disorders, *American Journal of Psychiatry*, 165:179–187.

Editorial (2018), Breaking the ice, *The Lancet Psychiatry*, 5:1.

Eells, E. (1991), *Probabilistic Causality*, Cambridge: Cambridge University Press.

Ehring, D. (1997), *Causation and Persistence*, Oxford: Oxford University Press.

Ehring, D. (2003), Physical Causation, *Mind*, 112:529–533.

Fair, D. (1979), Causation and the Flow of Energy, *Erkenntnis*, 14:219–250.

First, M. B. (2012), The National Institute of Mental Health Research Domain Criteria (RDoC) Project: Moving towards a Neuroscience-Based Diagnostic Classification in Psychiatry. In *Philosophical Issues in Psychiatry* II: *Nosology*. K. S. Kendler and J. Parnas, eds. Oxford: Oxford University Press.

Fodor, J. A. (1974), Special Sciences (or: The Disunity of Science as a Working Hypothesis), *Synthese*, 28:97–115.

Goldstein, R. Z. and Volkow, N. D. (2011), Dysfunction of the Prefrontal Cortex in Addiction: Neuroimaging Findings and Clinical Implications, *Nature Reviews Neuroscience*, 12:652–669.

Hall, N. (2004), Two Concepts of Causation. In *Causation and Counterfactuals*. J. Collins, N. Hall, and L. Paul, eds. Cambridge, MA: The MIT Press.

Halpern, J. Y. (2016), *Actual Causality*, Cambridge, MA: MIT Press.

Heyman, G. M. (2009), *Addiction: A Disorder of Choice*, Cambridge, MA: Harvard University Press.

Hoffman, G. A. and Zachar, P. (2017), RDoC's Metaphysical Assumptions: Problems and Promises. In *Extraordinary Science and Psychiatry: Responses to the Crisis in Mental Health Research*. J. Poland and Şerife Tekin, eds. Cambridge, MA: MIT Press.

Hyman, S. E. (2010), The Diagnosis of Mental Disorders: The Problem of Reification, *Annual Review of Clinical Psychology*, 6:155–179.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., and Wang, P. (2010), Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders, *American Journal of Psychiatry*, 167:748–751.

Kendler, K. S. (2011), Causal Thinking in Psychiatry: A Genetic and Manipulationist Perspective. In *Causality and Psychopathology: Finding the Determinants of Disorders*

and Their Cures. P. E. Shrout, K. M. Keyes, and K. Ornstein, eds. Oxford: Oxford University Press.
Kendler, K. S. and Campbell, J. (2009), Interventionist Causal Models in Psychiatry: Repositioning the Mind–Body Problem, *Psychological Medicine*, 39:881–887.
Kistler, M. (1998), Reducing Causality to Transmission, *Erkenntnis*, 48:1–24.
Kistler, M. (2006), *Causation and Laws of Nature*, London: Routledge.
Leshner, A. I. (1997), Addiction Is a Brain Disease, and It Matters," *Science*, 278:45–47.
Lewis, D. K. (1973), Causation, *Journal of Philosophy*, 70:556–567.
Lewis, M. D. (2015), *The Biology of Desire. Why Addiction Is Not a Disease*, New York: Public Affairs.
Lewis, M. D. (2017), Addiction and the Brain: Development, Not Disease, *Neuroethics*, 10:7–18.
Parnas, J. (2014), The RDoC Program: Psychiatry without Psyche?, *World Psychiatry*, 13:46–47.
Pearl, J. (2000), *Causality: Models, Reasoning, and Inference*, Cambridge: Cambridge University Press.
Pernu, T. K. (2011), Minding Matter: How Not to Argue for the Causal Efficacy of the Mental, *Reviews in the Neurosciences*, 22: 483–507.
Pernu, T. K. (2017), The Five Marks of the Mental, *Frontiers in Psychology*, 8:1084.
Pernu, T. K. (forthcoming), Mental Causation via Neuroprosthetics? A Critical Analysis, *Synthese*.
Polger, T. W., Shapiro, L. A. (2016), *The Multiple Realization Book*, Oxford: Oxford University Press.
Robins, L. N., Davis, D. H., and Goodwin, D. W. (1974), Drug Use by U.S. Army Enlisted Men in Vietnam: A Follow-Up on Their Return Home, *American Journal of Epidemiology*, 99: 235–249.
Robinson, T. E. and Berridge, K. C. (2000), The Psychology and Neurobiology of Addiction: An Incentive-Sensitization View, *Addiction*, 95:91–117.
Russell, B. A. W. (1948), *Human Knowledge: Its Scope and Limits*, London: George Allen & Unwin.
Russo, F. and Williamson, J. (2007), Interpreting Causality in the Health Sciences, *International Studies in the Philosophy of Science*, 21:157–170.
Salmon, W. (1984), *Scientific Explanation and the Causal Structure of the World*, Princeton, NJ: Princeton University Press.
Salmon, W. (1997), Causality and Explanation: A Reply to Two Critiques, *Philosophy of Science*, 64:461–477.
Schaffer, J. (2004), Causes Need not Be Physically Connected to Their Effects. In *Contemporary Debates in Philosophy of Science*. C. Hitchcock, ed. Oxford: Blackwell.
Shapiro, L. A. (2000), Multiple Realizations, *Journal of Philosophy*, 97:635–654.
Spirtes, P, Glymour, C., and Scheines, R. (2000), *Causation, Prediction and Search*, Cambridge, MA: MIT Press.

Suppes, P. (1970), *A Probabilistic Theory of Causality*, Amsterdam: North-Holland Publishing Company.

Szalavitz, M. (2016), *Unbroken Brain: A Revolutionary New Way of Understanding Addiction*, New York: St. Martin's Press.

Volkow, N. D., Koob, G. F., and McLellan, A. T. (2016), Neurobiologic Advances from the Brain Disease Model of Addiction, *New England Journal of Medicine*, 374:363–371.

Williamson, J. (2004), *Bayesian Nets and Causality: Philosophical and Computational Foundations*, Oxford: Oxford University Press.

Woodward, J. (2003), *Making Things Happen: A Theory of Causal Explanation*, New York: Oxford University Press.

Woodward, J. (2008), Cause and Explanation in Psychiatry: An Interventionist Perspective. In *Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology*. K. S. Kendler and J. Parnas, eds. Baltimore, MD: Johns Hopkins University Press.

12

Trauma-Informed Psychological Research

Ami Harbin

A number of areas of research in clinical psychology focus on understanding experiences of trauma and how individuals cope with and recover from them. In this chapter, I will consider how such research on trauma can raise complex important ethical, social, and political issues for researchers and clinicians to confront.

To make the case for the particular social and political issues that can arise when we look at traumas, I will begin the first section with a broad overview of areas of research on trauma in clinical psychology. I will survey some of these major areas, and offer an overview of different research methods and approaches employed. In the next section, I will introduce a concept which will help me draw out the social and political issues involved in these areas of investigation: *trauma-informed practice*. This is a term used to describe an approach now used in a number of professional fields, including in health care, education, and counseling, to indicate a central attentiveness to the effects of traumatic experiences on patients, students, or clients. After a summary of the central principles of trauma-informed practice, I will make the case in the last section for bringing a trauma-informed lens to research in clinical psychology. I will argue that this could help elucidate some of the social and political issues inherent to any context of research on trauma. More specifically, trauma-informed models might be useful when developing an approach to research on trauma in clinical psychology attentive to these important social and political issues.

Trauma research in clinical psychology

A great deal of research in clinical psychology addresses traumas and related experiences in some form. According to the American Psychological Association,

trauma is *an emotional response to a terrible event like an accident, rape, or natural disaster* (http://www.apa.org/topics/trauma/index.aspx). Typically, traumatic events cause emotional, physical, and psychological distress, and overwhelm individuals' sense of their own control, connection to others, and stability (Herman 1992: 33). Traumas can result from a number of kinds of life events, such as natural disasters and war.

Research in clinical psychology on trauma can employ a variety of approaches. A number of areas of quantitative research on traumas in clinical psychology employ scales and surveys of medium to large samples of participants to understand the experiences of those who have survived trauma, to evaluate the efficacy of treatment protocols, and to compare experiences of trauma across different groups. Qualitative research on traumas in clinical psychology tends to use interviews, focus groups, or other interactive measures with smaller samples of participants, which can allow for more in-depth understanding of the particular experiences of fewer people.

One area where traumas and post-traumatic recovery is investigated extensively is in the vast literature on post-traumatic stress disorder (PTSD). According to the *DSM*-5, common symptoms of PTSD include intrusive thoughts, reliving a traumatic event, experiencing physical and emotional reactions when reminded of the event, avoidance, loss of memory, diminished interest in everyday activities, emotional numbness, bleak outlook, insomnia, irritability, trouble concentrating, or hyper vigilance.[1] Researchers have investigated PTSD with a variety of goals in mind: for instance, to clearly understand particular symptoms of PTSD (e.g., Mellman and Hipolito 2006), to understand the neurobiology of PTSD (e.g., Krystal et al. 1989; Yehuda 1999), to understand trajectories of PTSD experiences over time (e.g., Santiago et al. 2013), to evaluate different screening methods (e.g., Brewin 2005; McDonald and Calhoun 2010), and to understand how different social identities or population membership may make a difference to one's experience of PTSD (e.g., Dyregrov and Yule 2006; Goff et al. 2007; Seedat and Stein 2000). In addition to these goals, much of the research on PTSD in clinical psychology focuses on understanding how PTSD can follow from particular kinds of events or experiences.

Perhaps the most common contexts understood to potentially trigger PTSD are those of military combat. Many of the earliest studies of PTSD focused on the experiences of soldiers and veterans following their involvement in wars (Kean et al. 1997). Some researchers have studied PTSD across different war eras (e.g., Magruder et al. 2009), while others have focused on particular military contexts (e.g., Blanchard 1990). Beyond military contexts, studies of

PTSD following illness have investigated, for example, how PTSD symptoms can follow experiences like heart attacks (Schaaf et al. 2013), severe burns (Lawrence and Fauerbach 2003), cancer (Pelcovitz et al. 1996), or treatment in intensive care units (Cuthbertson et al. 2004). Experiences of illness may be particularly traumatic when they are life-threatening, sudden, difficult to treat, and when they necessitate significant changes in individuals' everyday life. In conjunction with the other difficulties that accompany an illness, individuals may suffer symptoms of PTSD: intrusive thoughts, hypervigilance, avoidance, and so on. Similarly, the loss of a loved one and experience of bereavement can be traumatic and triggering of PTSD, particularly in cases where such loss is from "unnatural causes" (e.g., suicide or accident) (see Kaltman and Bonanno 2003; Zisook et al. 1998). Since loss can involve traumatic experiences for those left behind, in some cases, PTSD can be experienced alongside grief.

Some have suggested that a diagnosis of *Complex Post-Traumatic Stress Disorder* be used in cases where individuals have faced prolonged, repeated trauma (Herman 1992), such as in cases of long-term physical and/or sexual abuse (Roth et al. 1997), and often, though not always, following traumas occurring in childhood (Cloitre et al. 2009), though it is not currently a separate diagnostic category in the *DSM*-5 (American Psychiatric Association 2013). In cases of complex PTSD, survivors may experience difficulties with affect regulation, destructive behavior, dissociation, strained relationships with others, or loss of sustaining beliefs (Van der Kolk 2002). Here, too, there is now a great deal of research focusing on developing the most effective diagnostic tools, understanding of symptoms, and best practices for treatment (Cloitre et al. 2011).

When considering best therapeutic practices for responding to trauma, researchers in clinical psychology have investigated the effectiveness of a number of different approaches and strategies for intervention (Solomon et al. 1992). A variety of cognitive behavioral therapy approaches have been extensively considered (Foa et al. 1995; Foa and Rothbaum 2001), including exposure therapy, where survivors talk about or otherwise engage directly with their traumatic memories repeatedly in an effort to become less overwhelmed by or fearful of them (e.g., Gonçalves et al. 2012; Rauch et al. 2012), or eye-movement desensitization and reprocessing, where survivors talk or think about their traumatic memories while focusing on other stimuli, such as following a moving object with their eyes (Boudewyns and Hyer 1996). Sometimes in combination with cognitive behavioral therapy, survivors can be treated with psychotherapy in either individual or group settings (Bradley et al. 2005; Steenkamp et al. 2015). In some cases, survivors may receive psychopharmaceutical treatments,

with a number of drug options available to treat different symptoms (Asnis et al. 2004; Davidson 1992). Other treatment options approach recovery from trauma from a positive psychology perspective, and focus on survivors' capacities for resilience (Connor and Davidson 2003; Hoge et al. 2007; Lazarus 1966; Lazarus and Cohen 1977; Pietrzak et al. 2009; and Stewart and Yuen 2011), coping (Folkman and Lazarus 1988), hardiness (Gentry and Kobasa 1984; Kobasa 1979; Kobasa et al. 1982), stress-related growth (Park et al. 1996), and posttraumatic growth (Tedeschi and Calhoun 1996; 2004). And either as a stand-alone or complimentary practice, survivors may engage in holistic or integrative treatments, such as mindfulness, meditation, acupuncture, and relaxation techniques (Ventegodt et al. 2005; Wilson et al. 2001).

Before introducing the idea of trauma-informed practice, and considering how it may be helpful for highlighting some of the social and political issues inherent to any research on trauma in clinical psychology, I also want to note that philosophers have also engaged with clinical research on trauma from a variety of perspectives (e.g., see Brison 2002; Harbin 2016; Martin 2001; Nelson 2001; Potter 2006, 2015). Philosophical approaches can allow us to ask a number of important questions about clinical research on trauma. For instance, how have major philosophers since Plato understood and reflected on the position of trauma in human life? How can insights from neuroscience help strengthen psychological research on trauma? How do implicit assumptions about mind/brain dualism inform research on trauma? How do values shape psychological research on trauma? What unique responsibilities arise for clinicians in practicing with patients who have experienced trauma? What unique responsibilities arise for researchers in clinical psychology working with populations affected by trauma? How can we understand the autonomy and moral responsibility of patients who have experienced traumas? And how do social group-based harms (e.g., racism, sexism, heterosexism) shape clinical research on, diagnosis of, and treatment of traumas? Thus, the question I am asking here—*how can clinical research on trauma raise complex important ethical, social, and political issues for researchers and clinicians to confront?*—is just one of many possible philosophical questions about research on trauma in clinical psychology.

Now that we have considered a number of areas of research in clinical psychology on trauma, and philosophical questions they raise, I want to turn to a consideration of how trauma comes up in a number of other fields of medicine and professional work, and how some in those fields have prioritized understanding and responding to traumas as experienced by their client populations.

Trauma-informed practice

Trauma-informed approaches are currently employed in a number of fields where professionals work with individuals who have experienced trauma, and who are affected by those experiences in an ongoing way. Beyond their use in health care settings (Harris and Fallot 2001a, b, and c; Muskett 2014; Weissbecker and Clark 2007), trauma-informed practice has also been seen as important in first responder professions (e.g., fire fighters, emergency medical responders, and disaster response teams) (Ko et al. 2008), services for homeless people (Hopper et al. 2010), prisons (Miller and Najavits 2012) and in policing (Marans and Cohen 1993). They are also seen as valuable in all areas of work with children, including child welfare services (NCTSN Child Welfare Trauma Training Toolkit; Taylor et al. 2006), schools (Cognitive Behavioral Intervention for Trauma in Schools; Jaycox 2004; Stein et al. 2003), and juvenile justice facilities (Ford et al. 2007). Since traumatic experiences affect individuals who are accessing all kinds of professional services, from health care, to education, to counseling, and beyond, trauma-informed approaches are widely applicable.

According to the US Substance Abuse and Mental Health Services Administration, a trauma-informed approach "(1) Realizes the widespread impact of trauma … (2) Recognizes signs and symptoms of trauma in clients, families, staff, and others … (3) Responds by fully integrating knowledge about trauma into policies, procedures, and practices; and (4) Seeks to actively resist re-traumatization" (US Substance Abuse and Mental Health Services Administration). Trauma-informed approaches can be used in interactions with any clients, whether trauma itself is the reason they now seeking services or not. That is, *trauma-informed* approaches are meant to be more broadly applicable than *trauma-specific* interventions, which aim to directly address the impact or results of a particular trauma on an individual. As Butler and colleagues clarify:

> Trauma-informed services are informed about, and sensitive to, the potential for trauma-related issues to be present in patients, regardless of whether the issues are directly or obviously related to the presenting complaint or condition. Moreover, trauma-informed services are not designed to treat the sequelae of physical and sexual abuse or other traumatic experience. Trauma-specific services, in contrast, are designed expressly to treat the symptoms and syndromes related to current or past trauma. (Butler et al. 2011)

So, an educator could employ a trauma-informed approach in a classroom, even where the main goal is teaching, not identifying or treating a student's trauma,

and even when not everyone in the class has a history of trauma. A trauma-specific approach may be more evident in a health care or counseling setting where a clinician works with survivors of abuse and aims specifically to treat the symptoms the experience because of traumatic experiences in the past.

In some cases, trauma-informed approaches can be distinguished by the way they *conceptualize* trauma. In all such approaches, both the regularity and the impact of traumas are taken seriously. That is, trauma is understood to be an experience which many people have faced, and which is not rare, while at the same time being seen as an experience likely to have deep and sustained effects on a person's capacities for continuing in their life. As Harris and colleagues clarify:

> In a trauma-informed system, trauma is viewed not as a single discrete event but rather as a defining and organizing experience that forms the core of an individual's identity … in a trauma-informed system, practitioners assume that when trauma has occurred, it changes the rules of the game. An individual constructs a sense of self, a sense of others, and a belief about the world after trauma and abuse have occurred that incorporates and is in many cases based on the horrific event or events. (Harris and Fallot 2001c: 11–12)

In other words, trauma-informed approaches understand trauma as having holistic effects, which can augment an individual's identity, life course, and self-understanding, often in harmful ways.

Before a trauma-informed practitioner knows whether a client is affected by trauma, she can employ certain general practices with all clients that will be more likely to lead to successful relationships with any trauma-affected clients who may be present. As Raja and colleagues (2015) explain:

> Universal trauma precautions, which involve small changes to practice that can be employed with all patients, do not require providers to know whether or not a specific patient has a trauma history. These techniques may be particularly beneficial for establishing trust and rapport with trauma survivors. (Raja et al. 2015: 217)

Such universal precautions include, on Raja and colleagues' framework: patient-centered communication and care (e.g., asking every patient what could make them more comfortable during the appointment); understanding the health effects of trauma (e.g., understanding that maladaptive coping may be a result of trauma history); interprofessional collaboration (e.g., keeping referral material on trauma available in waiting room); understanding one's own history (e.g., of trauma) and reactions; and employing screening techniques (e.g., providing

staff with communication skills to perform trauma screenings) (Raja et al. 2015: 217, 219). To identify practices that exemplified trauma-informed care, Elliot and colleagues conducted research across nine comprehensive mental health and substance abuse service sites in the United States for women who had experienced trauma. On the basis of data from these sites, they identified ten principles of trauma-informed services: such services (1) recognize the impact of violence on development and coping strategies; (2) identify recovery from trauma as a primary goal; (3) employ an empowerment model; (4) strive to maximize a client's choices and control over their recovery; (5) are based in a relational collaboration; (6) create an atmosphere respectful of survivors' need for safety, respect, and acceptance; (7) emphasize clients' strengths, highlighting adaptations over symptoms and resilience over pathology; (8) aim to minimize the possibilities of retraumatization; (9) strive to understand each client in the context of their life experiences and cultural background; and (10) involve consumers in designing and evaluating services (Elliot et al. 2005: 465–469).

Trauma-informed approaches are interesting for many reasons, but the most important point to highlight about them for my purposes here is the way they grow out of a *non-individualistic awareness of the social and political dimensions of trauma*. That is, to understand both what trauma is and how trauma should be responded to, they think beyond just the individual service-user, to the broader community and society. In other words, they see trauma as a *social problem* (rather than an individual's problem), and seeing *responses to trauma as a community/social system's responsibility* (rather than an individual's responsibility alone). While each individual who has experienced trauma presents uniquely, having experienced both traumas and their effects in relevantly distinct ways, and always in the context of their own particular lives, we can notice that trauma-informed approaches treat trauma as a fundamentally social phenomenon, which does not require a response just from any one traumatized individual (e.g., actively seeking clinical treatment), but instead requires a response at the level of communities and social/structural systems (e.g., health care systems, education systems, criminal justice systems).

We see such approaches resisting treating traumas as an individual's *problem*, for example, through the awareness they demonstrate of the importance of not dismissing individuals' expressions of their experiences of traumas, and of not blaming victims for their own suffering. Instead, trauma-informed approaches take seriously survivors' expressions of their experiences and their needs, and employ a non-judgmental stance with survivors. They also show an awareness of how power dynamics in professional settings may keep clients from feeling

safe to disclose their past experiences, and recommend an approach that creates a safe space for everyone and prioritizes trust-building. Trauma-informed approaches aim to be aware of trauma as a feature of the world all of us share, and one which may disproportionately affect some members of society more than others—often those who are already vulnerable in other ways, because of gender, race, socioeconomic status, and so on. For instance, such approaches show an awareness of how certain traumas (e.g., sexual violence) can be very common and affect certain groups (women, and queer and trans communities) more than others. Such approaches then aim to develop practices sensitive to the needs of members of those groups; for example, in health contexts, some trauma-informed practices aim to minimize possible discomfort with certain medical exams by providing alternatives to having to undress fully during procedures (Raja et al. 2015).

Trauma-informed approaches also avoid approaching the task of responding to traumas as an individual survivor's *responsibility* through, for example, their insistence on the need for broad-based social supports to adequately respond to survivors' needs. For instance, Harris and Fallot (2001a) and the trauma committee of the Women, Co-occurring Disorders and Violence Study (funded by the Substance Abuse and Mental Health Services Administration) have focused on developing broader structural and systemic supports for people who have experienced trauma, understanding that there needed to be not just counseling targeted at processing the experience(s) of trauma, but also parenting support, health care, outreach, and resource coordination.

In sum, trauma-informed approaches to health care, social services, education, and other professions aim to develop principles and best practices for responding to individuals who may be experiencing the effects of trauma, and these approaches do so from a *non-individualistic awareness of the social and political dimensions of trauma,* understanding trauma as a *social problem* and *responses to trauma as a community/social system's responsibility*.

Trauma-informed psychological research

We saw in the section Trauma Research in Clinical Psychology that there are a number of areas of research on trauma and related experiences in clinical psychology—for instance, there is a great deal of research on post-traumatic stress disorders following experiences like military combat, illness, and grief, on complex post-traumatic stress disorder, and on best therapeutic practices for

responding to trauma. Such research may or may not be "trauma-informed"—it may or may not share the commitments of "trauma-informed practice" as we understood it in the section Trauma-Informed Practice. I now want to suggest that, by reflecting on research in clinical psychology through the lens of trauma-informed practice I have described, we can observe some important social and political issues that arise in research on trauma in clinical psychology, whether or not such issues are always apparent to those working in the field.

We might think that *research* in clinical psychology would be more distanced from social and political issues, and that it would be in the contexts of *clinical practice* or *policy making* that the true social and political issues would really arise. Crudely put, it might be tempting to think that scientific research is simply uncovering objective facts about the world, and that it is only when we need to decide what to do with those facts (i.e., how to allow clinical practice to be informed by them, or what policies to develop on the basis of them) that social and political complexities arise. On the contrary, I want to suggest (using the example of research on trauma) that the domain of research in clinical psychology can have significant social and political issues as well. Some of the same social and political issues that can arise in clinical practice or treatment for traumatic experiences can arise in different ways at the level of scientific research.

Social and political issues arise in research contexts from the very moment research questions are posed—why do we choose to focus on some questions rather than others? Whose perspectives are reflected in the questions we choose? Why are some research questions readily funded while others are not? And throughout the process of designing experiments, gathering data, interpreting data, establishing research conclusions, the values of the researchers and those that inform their broader research context shape what data is collected, how it is collected, and how it will be used (see Anderson 1995, 2004; Wylie 1996). Research in clinical psychology on trauma is no exception—which questions about trauma researchers focus on, how experiments on traumatic experiences and recovery are designed, how data is gathered and interpreted from survivors and others, and how conclusions are arrived at are thoroughly *value-laden* processes.

The area of research on trauma opens up particular social and political issues with respect to *how trauma and recovery will be understood and conceptualized* by research teams. The focus of trauma-informed practices described above can help elucidate this. Recall that trauma-informed approaches insisted on understanding trauma and recovery non-individualistically: they saw trauma

as a *social problem* (rather than an individual's problem), and seeing *responses to trauma as a community/social system's responsibility* (rather than an individual's responsibility alone). For research on trauma in clinical psychology, we can similarly ask: Does the research position trauma as an individual's problem, or as a social problem? And does the research position responding to trauma as an individual survivor's responsibility, or a community/social system's responsibility?

What might it mean for research on trauma in clinical psychology to position trauma as an individual's problem? Consider how particular research projects might make it seem as though traumatic experiences belong to individuals alone, rather than focusing on the effects of trauma within relationships, families, and communities. In some cases of research, the focus of the study of trauma's effects does not focus on how the social identity of the survivor might inform her experience of trauma, or may have made her more vulnerable to trauma in the first place—the demographics of research participants are not treated as an important focus of the study. What would it instead mean for such research to position trauma as a social problem? Such research could demonstrate awareness of trauma as a feature of the world all of us share, one which inevitably impacts an individual's loved ones and support networks, and one which may disproportionately affect some members of society more than others (i.e., as certain traumas [e.g., sexual violence, racialized violence] affect certain groups more than others).

What might it mean for research on trauma in clinical psychology to position the task of responding to trauma as an individual survivor's responsibility? Consider how research into the best therapeutic practices for responding to trauma may focus extensively on approaches to individual treatment (e.g., whether cognitive behavioral therapy or psychopharmaceutical or holistic approaches), while focusing very little on broader needs for social supports and community-based prevention. What would it mean for such research to position the task of responding to trauma as a community/social system's responsibility? Such research could, for example, focus more extensively on determining effective social supports for those who have experienced trauma, on investigating common barriers to accessing such supports where they exist, or on analyzing how clinical interventions can best integrate with broader-scale education and prevention programs (e.g., anti-sexual assault trainings). Of course, for such research to focus on better understanding possible societal responses to trauma would require an exploration of different methods or approaches to conducting studies. For instance, the research may need to become

more interdisciplinary, engaging scholars in public health, education, and other disciplines outside psychology. The research may also need to facilitate the involvement of service-users more thoroughly from the very beginning of study design, to ensure awareness of supports and barriers that the researchers may not otherwise anticipate. The necessary shifts in methods might not be obvious or easy to implement, but could be invaluable in developing trauma-informed psychological research.

Conclusion

Implicit within the language of "trauma-informed practice" is an understanding that practitioners should be informed about not only what trauma is and what effects it is likely to have, but also about where trauma has come from, and how it affects and must be responded to not only by individual survivors, but also their broader communities. As we have seen, bringing a trauma-informed lens to research in clinical psychology can help elucidate some of the social and political issues inherent to research on trauma in clinical psychology. Beyond highlighting their social and political dimensions, trauma-informed models might be useful for informing an approach to research on trauma in clinical psychology attentive to these important social and political issues.

Note

1 See the following PTSD symptom scale for further information: https://depts.washington.edu/hcsats/PDF/TF-%20CBT/pages/3%20Assessment/Standardized%20Measures/PSS-Adult.pdf.

References

American Psychiatric Association (2013), *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Arlington, VA: American Psychiatric Publishing.

Anderson, E. (1995), Knowledge, Human Interests, and Objectivity in Feminist Epistemology, *Philosophical Topics*, 23:27–58.

Anderson, E. (2004), Uses of Value Judgments in Science: A General Argument, with Lessons from a Case Study of Feminist Research on Divorce, *Hypatia*, 19:1–24.

Asnis, G. M., Kohn, S. R., Henderson, M., and Brown, N. L. (2004), SSRIs versus non-SSRIs in Post-Traumatic Stress Disorder, *Drugs*, 64:383–404.

Blanchard, E. B. (1990), Elevated Basal Levels of Cardiovascular Responses in Vietnam Veterans with PTSD: A Health Problem in the Making? *Journal of Anxiety Disorders*, 4:233–237.

Boudewyns, P. A. and Hyer, L. A. (1996), Eye Movement Desensitization and Reprocessing (EMDR) as Treatment for Post-Traumatic Stress Disorder (PTSD), *Clinical Psychology & Psychotherapy*, 3:185–195.

Bradley, R., Greene, J., Russ, E., Dutra, L., and Westen, D. (2005), A Multidimensional Meta-Analysis of Psychotherapy for PTSD, *American Journal of Psychiatry*, 162:214–227.

Brewin, C. R. (2005), Systematic Review of Screening Instruments for Adults at Risk of PTSD, *Journal of Traumatic Stress*, 18:53–62.

Brison, Susan J. (2002), *Aftermath: Violence and the Remaking of a Self*, Princeton, NJ: Princeton University Press.

Butler, L. D., Critelli, F. M., and Rinfrette, E. S. (2011), Trauma-Informed Care and Mental Health, *Directions in Psychiatry*, 31:197–208.

Cloitre, M., Courtois, C. A., Charuvastra, A., Carapezza, R., Stolbach, B. C., and Green, B. L. (2011), Treatment of Complex PTSD: Results of the ISTSS Expert Clinician Survey on Best Practices, *Journal of Traumatic Stress*, 24:615–627.

Cloitre, M., Stolbach, B. C., Herman, J. L., Kolk, B. V. D., Pynoos, R., Wang, J., and Petkova, E. (2009), A Developmental Approach to Complex PTSD: Childhood and Adult Cumulative Trauma as Predictors of Symptom Complexity, *Journal of Traumatic Stress*, 22:399–408.

Connor, K. M. and Davidson, J. R. T. (2003), Development of a New Resilience Scale: The Connor-Davidson Resilience Scale (CD-RISC), *Depression and Anxiety*, 18:76–82.

Cuthbertson, B. H., Hull, A., Strachan, M., and Scott, J. (2004), Post-Traumatic Stress Disorder after Critical Illness Requiring General Intensive Care, *Intensive Care Medicine*, 30:450–455.

Davidson, J. (1992), Drug Therapy of Post-Traumatic Stress Disorder, *The British Journal of Psychiatry*, 160:309–314.

Dyregrov, A. and Yule, W. (2006), A Review of PTSD in Children, *Child and Adolescent Mental Health*, 11:176–184.

Elliott, D. E., Bjelajac, P., Fallot, R. D., Markoff, L. S., and Reed, B. G. (2005), Trauma-Informed or Trauma-Denied: Principles and Implementation of Trauma-Informed Services for Women, *Journal of Community Psychology*, 33:461–477.

Foa, E. B. and Rothbaum, B. O. (2001), *Treating the Trauma of Rape: Cognitive-Behavioral Therapy for PTSD*, New York: Guilford Press.

Foa, E. B., Rothbaum, B. O., and Molnar, C. (1995), Cognitive-Behavioral Therapy of PTSD, pp. 483–494. In *Neurobiological and Clinical Consequences of Stress: From Normal Adaptation to PTSD*. M. J. Friedman, D. S. Charney, and A. Y. Deutsch, eds. Philadelphia, PA: Lippincott-Raven.

Folkman, S. and Lazarus, R. S. (1988), *The Ways of Coping Questionnaire*, Palo Alto: Consulting Psychologists Press.

Ford, J. D., Chapman, J. F., Hawke, J., and Albert, D. (2007), *Trauma among Youth in the Juvenile Justice System: Critical Issues and New Directions*, Delmar, NY: National Center for Mental Health and Juvenile Justice.

Gentry, W. D. and Kobasa, S. C. (1984), Social and Psychological Resources Mediating Stress-Illness Relationships in Humans, pp. 87–116. In *Handbook of Behavioral Medicine*. W. D. Gentry, ed. New York: Guilford Press.

Goff, A., Rose, E., Rose, S., and Purves, D. (2007), Does PTSD Occur in Sentenced Prison Populations? A Systematic Literature Review, *Criminal Behaviour and Mental Health*, 17:152–162. doi:10.1002/cbm.653.

Gonçalves, R., Pedrozo, A. L., Coutinho, E. S. F., Figueira, I., and Ventura, P. (2012), Efficacy of Virtual Reality Exposure Therapy in the Treatment of PTSD: A Systematic Review, *PloS One*, 7:e48469.

Harbin, A. (2016), *Disorientation and Moral Life*, New York: Oxford University Press.

Harris, M. and Fallot, R. D. (2001a), Envisioning a Trauma-Informed Service System: A Vital Paradigm Shift, *New Directions for Mental Health Services*, 89:3–22.

Harris, M. and Fallot, R. D. (2001b), Trauma-Informed Inpatient Services, *New Directions for Mental Health Services*, 89:33–46.

Harris, M. and Fallot, R. D. (2001c), *Using Trauma Theory to Design Service Systems*, San Francisco, CA: Jossey-Bass.

Herman, J. L. (1992), Complex PTSD: A Syndrome in Survivors of Prolonged and Repeated Trauma, *Journal of Traumatic Stress*, 5:377–391.

Hoge, E. A., Austin, E. D., and Pollack, M. H. (2007), Resilience: Research Evidence and Conceptual Considerations for Posttraumatic Stress Disorder, *Depression and Anxiety*, 24:139–152.

Hopper, E. K., Bassuk, E. L., and Olivet, J. (2010), Shelter from the Storm: Trauma-Informed Care in Homelessness Services Settings, *The Open Health Services and Policy Journal*, 3:80–100.

Jaycox, L. (2004), *Cognitive Behavioral Intervention for Trauma in Schools*, Longmont, CO: Sopris West Educational Services.

Kaltman, S., and Bonanno, G. A. (2003), Trauma and Bereavement: Examining the Impact of Sudden and Violent Deaths, *Journal of Anxiety Disorders*, 17:131–147.

Keane, T. M., Newman, E., and Orsillo, S. M. (1997), *Assessment of Military-Related Posttraumatic Stress Disorder*, New York: Guilford Press.

Ko, S. J., Ford, J. D., Kassam-Adams, N., Berkowitz, S. J., Wilson, C., Wong, M., Brymer, M. J., and Layne, C. M. (2008), Creating Trauma-Informed Systems: Child Welfare, Education, First Responders, Health Care, Juvenile Justice, *Professional Psychology: Research and Practice*, 39:396–404.

Kobasa S. C. (1979), Stressful Life Events, Personality, and Health: An Inquiry into Hardiness, *Journal of Personality and Social Psychology*, 37:1–11.

Kobasa S. C., Maddi S. R., and Kahn S. (1982), Hardiness and Health: A Prospective Study, *Journal of Personality and Social Psychology*, 42:168–177.

Krystal, J. H., Kosten, T. R., Southwick, S., Mason, J. W., Perry, B. D., and Giller, E. L. (1989), Neurobiological Aspects of PTSD: Review of Clinical and Preclinical Studies, *Behavior Therapy*, 20:177–198.

Lawrence, J. W. and Fauerbach, J. A. (2003), Personality, Coping, Chronic Stress, Social Support and PTSD Symptoms among Adult Burn Survivors: A Path Analysis, *Journal of Burn Care & Research*, 24:63–72.

Lazarus, R. S. (1966), *Psychological Stress and Coping*, New York: McGraw–Hill.

Lazarus, R. S. and Cohen, J. B. (1977), Environmental Stress. In *Human Behavior and Environment*, vol. 2. I. Altman and J. F. Wohlwill, eds. New York: Plenum.

Magruder, K. M. and Yeager, D. E. (2009), The Prevalence of PTSD across War Eras and the Effect of Deployment on PTSD: A Systematic Review and Meta-Analysis, *Psychiatric Annals*, 39:778.

Marans, S. and Cohen, D. (1993), Children and Inner-City Violence: Strategies for Intervention, pp. 281–301. In *Psychological Effects of War and Violence on Children*. L. Leavitt and N. Fox, eds. Hillsdale, NJ: Erlbaum.

Martin, N. (2001), Feminist Bioethics and Psychiatry, *Journal of Medicine and Philosophy*, 26:431–441.

McDonald, S. D. and Calhoun, P. S. (2010), The Diagnostic Accuracy of the PTSD Checklist: A Critical Review, *Clinical Psychology Review*, 30:976–987.

Mellman, T. A. and Hipolito, M. M. S. (2006). Sleep Disturbances in the Aftermath of Trauma and Posttraumatic Stress Disorder, *CNS Spectrums*, 11:611–615.

Miller, N. A. and Najavits, L. M. (2012), Creating Trauma-Informed Correctional Care: A Balance of Goals and Environment, *European Journal of Psychotraumatology*, 3:17246.

Muskett, C. (2014), Trauma-Informed Care in Inpatient Mental Health Settings: A Review of the Literature, *International Journal of Mental Health Nursing*, 23:51–59.

Nelson, H. L. (2001), *Damaged Identities, Narrative Repair*, Ithaca, NY: Cornell University Press.

Park, C., Cohen, L. and Murch, R. (1996), Assessment and Prediction of Stress-Related Growth, *Journal of Personality*, 64(1):71–105.

Pelcovitz, D., Goldenberg, B., Kaplan, S., Weinblatt, M., Mandel, F., Meyers, B., and Vinciguerra, V. (1996), Posttraumatic Stress Disorder in Mothers of Pediatric Cancer Survivors, *Psychosomatics*, 37:116–126.

Pietrzak R. H., Johnson, D. C., Goldstein, M. B., Malley, J. C., and Southwick, S. M. (2009), Psychological Resilience and Postdeployment Social Support Protect against Traumatic Stress and Depressive Symptoms in Soldiers Returning from Operations Enduring Freedom and Iraqi Freedom, *Depression and Anxiety*, 26:745–751.

Potter, N. N. (2006), *Trauma, Truth, and Reconciliation: Healing Damaged Relationships*, New York: Oxford University Press.

Potter, N. N. (2015), Reflections on Things We Don't Want to Think About: Intersections of Colonialism, Transgenerational Trauma, and Oppression within Psychiatry, *Journal of Ethics in Mental Health*. URL= http://www.jemh.ca/issues/

open/documents/JEMH_Open-Volume_Article_Theme_Colonization_Reflections_June2015.pdf.

Raja, S., Hasnain, M., Hoersch, M., Gove-Yin, S., and Rajagopalan, C. (2015), Trauma Informed Care in Medicine Current Knowledge and Future Research Directions, *Family and Community Health*, 38:216–226.

Rauch, S. A., Eftekhari, A., and Ruzek, J. I. (2012), Review of Exposure Therapy: A Gold Standard for PTSD Treatment, *Journal of Rehabilitation Research and Development*, 49:679–688.

Roth, S., Newman, E. Pelcovitz, D., Van Der Kolk, B., and Mandel, F. S. (1997), Complex PTSD in Victims Exposed to Sexual and Physical Abuse: Results from the DSM-IV Field Trial for Posttraumatic Stress Disorder, *Journal of Traumatic Stress*, 10:539–555.

Santiago, P. N., Ursano, R. J., Gray, C. L., Pynoos, R. S., Spiegel, D., Lewis-Fernandez, R., and Fullerton, C. S. (2013), A Systematic Review of PTSD Prevalence and Trajectories in DSM-5 Defined Trauma Exposed Populations: Intentional and Non-Intentional Traumatic Events, *PloS One*, 8:e59236.

Schaaf, K. P. W., Artman, L. K., Peberdy, M. A., Walker, W. C., Ornato, J. P., Gossip, M. R., and Kreutzer, J. S. (2013), Anxiety, Depression, and PTSD Following Cardiac Arrest: A Systematic Review of the Literature, *Resuscitation*, 84:873–877.

Seedat, S. and Stein, D. J. (2000), Trauma and Post-Traumatic Stress Disorder in Women: A Review, *International Clinical Psychopharmacology*, 15:S25-S33.

Solomon, S., Gerrity, E., and Muff, A. (1992), Efficacy of Treatments for Posttraumatic Stress Disorder: An Empirical Review, *JAMA*, 268:633–638.

Steenkamp, M. M., Litz, B. T., Hoge, C. W., and Marmar, C. R. (2015), Psychotherapy for Military-Related PTSD: A Review of Randomized Clinical Trials, *JAMA*, 314:489–500.

Stewart, D. and Yuen, T. (2011), A Systematic Review of Resilience in the Physically Ill, *Psychosomatics*, 52:199–209.

Stein, B. D., Jaycox, L. H., Kataoka, S. H., Wong, M., Tu, W., Elliot, M. N., et al. (2003), A Mental Health Intervention for Schoolchildren Exposed to Violence, *Journal of the American Medical Association*, 290:603–611.

Taylor, N., Wilson, C., and Igelman, R. (2006), In Pursuit of a More Trauma-Informed Child Welfare System, *APSAC Advisor*, 18(2):4–9.

Tedeschi, R. G. and Calhoun, L. G. (1996), The Posttraumatic Growth Inventory: Measuring the Positive Legacy of Trauma, *Journal of Traumatic Stress*, 9:455–471.

Tedeschi, R. G. and Calhoun, L. G. (2004), Posttraumatic Growth: Conceptual Foundations and Empirical Evidence, *Psychological Inquiry*, 15:1–18.

Van Der Kolk, B. A. (2002), The Assessment and Treatment of Complex PTSD, pp. 127–156. In *Treating Trauma Survivors with PTSD*. R. Yehuda, ed. Washington, DC: American Psychiatric Publishing.

Ventegodt, S., Kandel, I., Neikrug, S., and Merric, J. (2005), Clinical Holistic Medicine: Holistic Treatment of Rape and Incest Trauma, *The Scientific World Journal*, 5:288–297.

Weissbecker, I. and Clark, C. (2007), The Impact of Violence and Abuse on Women's Physical Health: Can Trauma-Informed Treatment Make a Difference?" *Journal of Community Psychology*, 35:909–923.

Wilson, J. P., Friedman, M. J., and Lindy, J. D. (2001), A Holistic, Organismic Approach to Healing Trauma and PTSD, pp. 28–566. In *Treating Psychological Trauma and PTSD*. J. P. Wilson, M. J. Friedman, and J. D. Lindy, eds. New York: The Guilford Press.

Wylie, A. (1996), The Constitution of Archaeological Evidence: Gender Politics and Science. In *The Disunity of Science*. P. Galison and D. Stump, eds. Palo Alto, CA: Stanford University Press.

Yehuda, R. (1999), Linking the Neuroendocrinology of Post-Traumatic Stress Disorder with Recent Neuroanatomic Findings, *Seminars in Clinical Neuropsychiatry*, 4:256–265.

Zisook, S., Chentsova-Dutton, Y., and Shuchter, S. R. (1998), PTSD following Bereavement, *Annals of Clinical Psychiatry*, 10:157–163.

Study Questions

1. *Are observations about mental disorders theory-laden?*
2. *What kind of science is psychiatry?*
3. *What roles do values play in the scientific research on mental disorders?*
4. *What is causal explanation in psychiatry?*
5. *Are mental disorders real or socially constructed?*

Part Five

Ethics and Psychiatry

13

Informed Consent in Psychiatry: Philosophical and Legal Issues

Claire Pouncey and Jon F. Merz

Introduction

Informed consent, a central tenet of biomedical treatment and research, is the autonomous, voluntary act of authorizing one's own medical treatment with awareness of one's illness, need for treatment, treatment options, and the risks and potential benefits of each, considered in light of one's unique situation and personal values. Despite its centrality, informed consent is a difficult goal to achieve, even in the best of circumstances. Paradoxically, informed consent often is further complicated by limitations imposed by the illness treatment is meant to alleviate. This is particularly true in psychiatry, where the limitations treatment is meant to address, and the settings in which treatment is offered, make informed consent both conceptually and practically challenging.

Our aim is to give an overview of the philosophical issues in informed consent that psychiatrists face in regular practice. In this chapter, we discuss the legal, ethical, and epistemological issues that ground the doctrine of informed consent in medicine and psychiatry. We focus primarily on informed consent for treatment, with consideration of informed consent for research participation at the end. We have two goals. First, we present informed consent as an ongoing, collaborative procedure between care providers and patients that maximizes patient self-determination and minimizes professional coercion and overreach. Second, we illuminate the philosophical issues surrounding informed consent that psychiatrists face regularly, both in the role of consultation psychiatrists and as care providers in different psychiatric treatment settings.

The ethics and epistemology of informed consent

Following Faden and Beauchamp (1986), we define informed consent for medical treatment as the autonomous, voluntary process of authorizing one's own medical treatment, with awareness of one's illness, need for treatment, treatment options, and the risks and potential benefits of each, considered in light of one's personal situation and values. In a clinical setting, informed consent supports patients' right to self-determination, the right to decide what course of treatment, if any, they wish to pursue. Treatments are offered and performed to alleviate pain and suffering, or to extend life. Informed consent promotes the ability of ill and injured patients to make reasoned treatment decisions that they, in consult with their treatment team, family, other health care providers, and perhaps close friends, believe will be best for them. Informed consent thus empowers patients to meaningfully contribute to decisions that affect their health care and ultimate well-being. The ethical doctrine of informed consent mandates that practicing physicians involve their patients in decision making as much as possible, to the extent patients wish to play an active role (Meisel 1996).

Such self-determination presupposes that patients have full moral and epistemic agency. Agency is the ability to act on one's own behalf. Moral agency is the ability to act volitionally and deliberately as a member of a moral community in accordance with shared ethical values and principles that govern (1) what is right or wrong to do, and (2) how transgressors are to be held accountable. Two familiar philosophical examples of theoretical moral communities are Hobbes's view of moral agents living in accordance with the rules of a tacit social contract (Hobbes 1994) and Kant's description of moral agents mutually participating in a moral community according to recognized ethical obligations (Kant 1993). However, moral agents may comprise much more informal and local communities as well. A moral community may be any group that tacitly or explicitly creates rules to govern interpersonal behavior in that society, such as the code of ethics of a school or a professional organization. The ethics of informed consent establishes a moral community comprised of doctors and patients. By making patients responsible for their own treatment choices, the doctrine of informed consent makes patients full moral agents, full participants in choosing what a physician may or may not do to that person's body.

Informed consent also presupposes epistemic agency. In philosophy, epistemology is the study of knowledge, and includes questions about what we accept as knowledge and the bases for doing so. Just as a moral community can be defined by shared values and rules of interpersonal conduct, an epistemic

community can be defined as a community of knowers, epistemic agents with shared views of how knowledge is created and perpetuated, why some beliefs and not others may be considered knowledge, what constitutes expertise rather than commonplace knowledge, and how to utilize knowledge in the process of making decisions, from mundane daily choices to more significant deliberations in light of overall life values. Epistemic agency is the ability to be a full cognitive and intellectual participant in a community of knowers.

Although informed consent is often considered an ethical doctrine, the epistemological issues have always been central. Yet the epistemology of informed consent is seldom discussed explicitly in the medical, bioethics, or philosophy literatures. The traditional view of informed consent names five fundamental components: (1) adequate *disclosure* of medical information by the physician, (2) the patient's *capacity* for autonomous decision making, (3) the patient's *understanding* of medical information and *reasoning* about her own values and best interests, (4) the *voluntariness* of the patient's choice, and (5) the patient's *authorization* for treatment (Appelbaum and Gutheil 2007; Beauchamp and Childress 2001; Faden and Beauchamp 1986; National Commission 1979). We find that listing the elements of informed consent this way obscures the relationships among them. Understanding informed consent in terms of both moral and epistemic agency better captures the relationships among the components and communicates the ideals behind informed consent doctrine and practice, as we will go on to discuss. The interrelatedness of the epistemological concepts of disclosure, capacity, understanding, and reasoning allows us to appreciate how philosophically complex the doctrine of informed consent is in medicine generally, but especially in psychiatric care. Informed consent is not only a moral tenet about how doctors treat patients as members of a common moral community, but is also a guideline for doctors to bring patients into a shared epistemic community. Although informed consent is usually considered the purview of medical ethics, epistemological concerns are central to ordinary applications of the doctrine.

Before the doctrine of informed consent took hold, physicians' authority and expertise subsumed patients' moral and epistemic agency. Patients and doctors alike subscribed to a "doctor knows best" or "trust me, I'm the doctor" mentality. The role of the patient was to be compliant; the role of the doctor was to apply expert knowledge to treat the ill. However, physicians did not always provide prognoses, or even diagnoses, and they did not always discuss treatment options and the expected risks and potential benefits inherent therein with patients. Physicians selectively disclosed details about the relevant illness, the need for

treatment, the range of available treatments and their potential benefits and harms, and the predicted consequences of accepting or refusing treatment (Katz 1984). These omissions denied patients full agency in making health care decisions. They denied epistemic agency by withholding medical knowledge necessary for patients to make their own treatment decisions, and they denied moral agency because without that knowledge, patients could not make truly self-determined decisions about their overall interests.

The doctrine of informed consent is a dictum for doctors to become educators and collaborators about their treatment recommendations. Informed consent requires treating patients as full moral agents capable of making their own life choices according to their own aims and values. Informed consent also requires treating patients as full epistemic agents capable of acquiring, utilizing, and applying pertinent medical knowledge to serve their own aims. In the clinical setting, this requires not only the disclosure of medical information regarding a person's diagnosis, treatment options, and their attendant risks and benefits, but also requires disclosure in a manner that a fellow epistemic agent without medical expertise can adequately comprehend. The doctrine of informed consent presupposes that patients have the ability to appreciate their own personal and medical needs, the health provider's rationale for recommending treatment, and the pros and cons of both receiving and refusing treatment. In other words, informed consent presupposes patients' abilities to draw their own rational conclusions and make unrestricted choices about whether to authorize a recommended treatment or not.

The law and informed consent

The philosophy and law of informed consent evolved hand-in-hand. As a general matter, the law has not, with one notable exception (Helling v Carey 1974), led the way, instead deferring to medical practices to establish legal standards of care.[1] This is true for informed consent in both clinical settings and research, though the path to legal standards in these two domains varied greatly. More critically, for practicing physicians and researchers, what the law requires in treatment and research settings also differs.

The law has required consent of patients to any physical touching of their body by their doctor for well over a century (Schloendorff v. The Society of the New York Hospital 1914). Informed consent evolved from the view that consent must be based on adequate disclosure of information, and failure to adequately

inform a patient will vitiate the consent given. The first legal case to use the term "informed consent" was *Salgo v. LeLand Stanford Jr. Board of Trustees* (1957). Through the 1960s and '70s, the courts in numerous states recognized a cause of legal action for lack of informed consent. Critically, the doctrine of informed consent provides compensation for the foreseeable risks of care that occur regardless of the quality (and legal sufficiency) of the care rendered.

One of the early courts to recognize the doctrine of informed consent wrestled with the problem posed by the logical conclusion that if a patient is not adequately informed, then the consent is not valid, resulting in a battery (Natanson v. Kline 1961). Battery is the intentional, nonconsensual touching of another, and gives rise to potential civil liability for damages caused by the touching, incidental damages for the dignitary insult, and potentially punitive damages to deter such behaviors in the future. Malpractice insurance typically excludes coverage for such intentional acts, leaving physicians exposed. Moreover, the law seeks to promote physician disclosures of risks, potential benefits, and treatment alternatives to patients, and the legal inquiry as to whether a physician did so is more akin to negligence, particularly when the physician did secure a patient's expression of consent, and only later may be adjudged to be based upon a legally inadequate disclosure. Thus, most courts that adopted the doctrine focused on the physician's behavior, characterizing the failure to adequately inform to be actionable negligence or malpractice.

A second problem that the courts faced was establishing causation. Liability for professional negligence requires that four elements be established: (1) a duty, (2) breach of that duty, (3) injury for which monetary damages can be awarded, and (4) causation of that injury. In informed consent, a doctor has a legal duty to adequately inform patients, and failure to so inform comprises a breach of that duty. The injury complained of must have arisen from the care provided under the deficient informed consent and, but for the lack of disclosure, the injury would not have occurred. Causation in this case is bifurcated: the injury must have been physically caused by the medical care rendered ("injury causation"), and, in addition, the decision the patient made would likely have been different if a more complete disclosure was made ("decision causation") (Meisel and Kabnick 1980). An analysis of published informed consent cases strongly suggested that decision causation acts to limit the disclosure obligation, because patients generally will go along with their doctor's recommendations regardless of what they are told (Merz 1991). This problem is exacerbated by courts' focus on disclosure of the known risk that actually transpired in the case in front of them, ignoring whether other

nondisclosed risks—which did not come to fruition—would have altered the patient's decision (Merz et al. 1993).

A third problem for courts was determining by what standard to judge whether a patient would have made a different decision. Negligence law is typically focused on the behaviors of the fictional reasonable person, and it would be unfair to hold physicians liable for truly idiosyncratic decisions that aggrieved patients would assert that they would have made. Thus, US courts uniformly held that liability could only be found if, along with the other elements, the jury determined that a reasonable person in the patient-plaintiff's position would have chosen differently if a more complete disclosure had been made (Canterbury v. Spence 1972). New Zealand and Canada, notably, utilize a subjective patient standard; in reality, there may be little difference, because an aggrieved patient would have to convince a jury that they really would have made a different decision, which could be difficult to do.

A fourth problem for courts was determining how to assess the adequacy of disclosure. By what standard is the law to assess the sufficiency of what was told to the patient, the way it was told, when, and how much opportunity there was for the patient to assimilate and consider the information in making a decision? Many courts followed the general rule of negligence law, requiring health care providers to disclose the information that other providers would disclose under the same circumstances. Many other courts, in contrast, applied a "prudent patient" standard, requiring physicians to disclose that information which a reasonably prudent patient would wish to know, or the information that would be "material" to a reasonable person in the patient's position. Several courts required both (Cobbs v. Grant 1972).

In response to the broad adoption of informed consent as a legal basis for physician malpractice liability through the 1970s, a number of states enacted informed consent legislation in the 1980s to minimize that liability. Many of these statutes reacted to what they perceived as unrealistic burdens on physicians by severely limiting physicians' disclosure obligations, typically requiring a standardized disclosure that could easily be satisfied by health care providers, thus abandoning the lofty ethical goals of promoting physician–patient discussions, information exchange, and shared decision making.

There are three take-away lessons from this review. First, the law is a blunt instrument that serves many purposes; practicing physicians should be aware of the informed consent requirements in the states in which they practice, but they should not look to the law to dictate how to behave.

Second, practicing physicians should seek to involve their patients in decision making as much as possible, to the extent patients wish to play an active role (Meisel and Kuczewski 1996). Neither law nor ethics dictate that patients must decide, but it is for the physician and the patient to work out a reasonable way for the patient to share the making of and responsibility for treatment decisions, to consider her options and bring her own values, hopes, experiences, and expectations to bear.

Third, it is important to distinguish two different senses of "informed consent" (Faden and Beauchamp 1986), which we will call "ethical" and "administrative" informed consent. Medical trainees are familiar with administrative informed consent. Treatment team members in various medical specialties talk about whether a patient awaiting a procedure has been "consented" yet, and whether "the consent is on the chart," which makes informed consent sound like a check box on a legal-administrative to-do list rather than a collaborative, educational process. Administrative informed consent reflects the legal reinforcement of the ethical doctrine of informed consent. Ironically, however, the legal mechanisms for enforcing the informed consent process skew perceptions of informed consent from an ongoing, collaborative, educational process to a documented product. Administrative informed consent follows the letter but not the intent of informed consent doctrine. To protect themselves from malpractice claims, physicians and hospitals document patients' consent to hospitalization, to medications, to studies, and to procedures. Such documentation, however, records in general terms that a conversation has taken place and that the patient agrees to care. This administrative informed consent, insofar as its aim is to protect practitioners and institutions from malpractice lawsuits, creates an adversarial rather a collaborative relationship with patients (Tauber 2005). From here on, our discussion will consider informed consent as an ongoing, collaborative process rather than as a documented product or an administrative formality.

Psychosomatic psychiatry and capacity determinations

We mentioned that informed consent is traditionally taken to have five components: disclosure, capacity, reasoning and understanding, voluntariness, and authorization. It is helpful to think of two of these as *pre*conditions for informed consent. The first is disclosure: we have seen that the law requires physician disclosure of relevant medical information, because informed consent

is impossible if the patient is ignorant of her medical condition, the proposed treatment, and the expected risks and benefits of both treatment and treatment alternatives. In philosophical terms, the physician must bring the patient into the shared epistemic community by educating the patient about her condition, the available treatments, and their risks and potential benefits. The second precondition is capacity: the need for informed consent is vitiated if the patient lacks capacity to understand the disclosed information and give voluntary, informed authorization.

The literature on informed consent sometimes uses the terms "capacity" and "competence" interchangeably. We respect the convention that physicians make capacity assessments, while courts make competency determinations. There are two differences. The first is that a legal determination of incompetence is a global determination that a person is not recognized as a *legal* agent, someone who can make her own legal decisions about making a will, standing trial, living alone, etc. Children lack legal agency until an age of majority, usually eighteen, and until that time parents or other legal guardians are typically their medical decision makers. Persons with neurocognitive or intellectual impairments may also be found to be legally incompetent and assigned guardians who would have legal authority to consent on the patient's behalf. Capacity becomes an issue for persons who are assumed to be legal and epistemic agents, but whose epistemic agency seems to be compromised.

The second difference between legal competence and capacity is that capacity is determined with respect to specific tasks. Because of her height, someone might have the capacity to change a lightbulb on a table lamp but not a ceiling lamp. Because of education level, someone might have the capacity to do simple arithmetic but not differential equations. When we speak of capacity with respect to informed consent for treatment decisions, the specific question is whether a person can, at a given time, comprehend her own medical situation and the information provided by her doctor, appreciate the circumstances she is in, reason about her options in light of her own life circumstances and values, and authorize or refuse the recommended treatment.

Capacity thus raises a logical problem in that the definition of capacity is circular: informed consent is partly defined in terms of capacity (as one of the five constitutive elements), and capacity is defined as the ability to give informed consent for a given medical procedure. For instance, "In a medical context, capacity refers to the ability to utilize information about an illness and proposed treatment options to make a choice that is congruent with one's own values and preferences" (Karlawish 2016). Or for another example, capacity is

defined as the sufficient ability to make a meaningful treatment decision given the circumstances the patient faces (Appelbaum and Gutheil 2007: 196). Most authors do not comment on this circularity or treat it as problematic. Since the specific task at hand is to make an informed, autonomous medical treatment decision, the capacity in question refers to that task and not, say, the capacity to rotate one's tires or make a last will and testament. Yet technically, the definition is circular. This circular definition of capacity makes it impossible to define clear standards for ascertaining the degree to which a particular person has capacity to make a given treatment decision, or in other words, whether that person's epistemic agency is limited. This logical challenge becomes the purview of psychosomatic psychiatry in the "capacity consultation."

In most health care settings, epistemic agency is assumed. Capacity assessments are tacit and informal, and occur as part of everyday medical examinations. Treatment providers generally do not question whether a patient is capable of comprehending the nature of her illness and the treatment options available, nor do they require the patient to justify a treatment decision or demonstrate the reasoning by which she reached it. In hospitals and other acute care settings, however, treating physicians often call upon psychiatrists to determine whether a patient has the capacity to authorize or—more frequently—refuse treatment. Although all physicians can make capacity determinations, psychiatrists are called upon for formal capacity assessments to ascertain whether a patient is capable of making a certain treatment decision at a particular point in time.

Although physicians typically assume their patients' epistemic agency, the literature on health literacy challenges this assumption. Physicians tend to assume that patients have coherent world views, rational and consistent ways of making decisions, and cognitive methods of predicting how various treatment options align with their life goals and values. They assume that patients understand relative risks in the face of medical uncertainty. Care providers often assume that patients understand the technical medical information they are given and use it in their decision making. These assumptions ignore documented problems with health literacy, the ability to obtain, process, and comprehend basic health information and health services. The US Department of Education reports that only 12 percent of English-speaking adults are proficient in understanding health information and health services (Kutner et al. 2006). For informed consent to be authentic, however, patients actually must be able to understand the medical information provided, which means that physicians must take the time to explain illnesses and treatments, possibly more than once, and to answer patients' questions as needed to make them fellow knowers. They must

be able to manipulate that information in light of their own medical and life circumstances, in accordance with their own values, and do so in a setting in which treatment participation can be considered truly voluntary.

The medical condition requiring treatment can also limit epistemic agency. Epistemic expectations are frequently unmet in acute medical care settings simply by nature of the limitations illness creates. Pain, infection, metabolic changes, physical trauma, neurocognitive, psychiatric, and other symptoms can affect a patient's attention to and comprehension of medical information. Those symptoms may also limit that person's ability to appreciate how the information disclosed is relevant to their personal circumstances. It is difficult to become a knower when illness or trauma limits one's ability to think about anything else.

Finally, the medical treatment setting itself may limit a patient's epistemic agency. Patients facing medical treatment decisions are often afraid and distracted by the sounds and events of an acute care setting, and they may lack the social support they need to process information thoroughly. They may be intimidated by real or perceived professional authority or time constraints, or simply feel uncertain about having the ultimate say over their course of treatment.

For all of these reasons and more, epistemic agency can be problematic in acute medical settings. When a patient agrees to a treatment without fully understanding what it is, whether it is absolutely necessary, or what its risks and benefits might be, that patient does not give informed consent, but merely *assents* to treatment. That is, he does not refuse treatment, but he does not voluntarily and knowingly authorize it in the manner informed consent requires (Faden and Beauchamp 1986). And without voluntary authorization, there is also a failure of moral agency, a failure to act volitionally in accordance with the value of self-determination shared by the moral community. This failure of epistemic and moral agency entails a failure of informed consent. Similarly, when a patient refuses a treatment without full understanding, the failure of epistemic and moral agency entail an uninformed refusal.

Enter the psychiatrist. Treating physicians in acute care settings usually call for psychiatric assessments of a patient's capacity—"capacity consultations"—when the patient refuses a recommended treatment. As we noted above, as long as the patient agrees with the physician's recommendations, little scrutiny is given to patient capacity, even when the agreement constitutes assent, an explicit failure of informed consent. When the patient is legally incompetent to make her own treatment decisions, either because of youth or because of legally determined, widespread limitations in decision-making abilities, a substitute decision maker such as a family member or legal guardian will be called upon to make treatment

decisions. Treating physicians ask a psychiatrist to consult when a patient refuses a treatment or diagnostic procedure the treating doctor believes is medically necessary. In other words, the treating physician questions a person's agency when the patient acts like a moral and epistemic equal, but reaches a different conclusion than the physician about what is in the patient's best interests. To a physician, if a patient presents for medical treatment and then refuses it after the doctor explains her recommendations, there appears to be an aspect of the treatment the patient does not or cannot understand, even if the disagreement is about values rather than facts. If the recommended procedure is very urgent, such as a cardiac catheterization after a myocardial infarction, surgery after a trauma, or intravenous antibiotics to treat sepsis, the treating physician is likely to consult a psychiatric colleague for a formal capacity evaluation.

Capacity evaluations must be made with respect to a specific treatment decision at a specific time. When a physician requests a "capacity consultation" from a psychiatrist, the question to be answered should be phrased precisely: "Does Ms. W. have capacity to refuse the cardiac pacemaker placement scheduled for tomorrow morning?" The physician should also inform the consultant of the content of the medical disclosure and the ensuing dialogue (ideally, by documenting it in the patient's medical record). The psychiatrist's responsibility is to ascertain whether Ms. W. understands the medical information *already disclosed* to her by her doctor, and whether she can utilize that information in accordance with her own beliefs and values to refuse or authorize the treatment. The consultation cannot begin if the treatment team has not fulfilled the disclosure precondition. Although an eventual determination that Ms. W. lacks capacity retroactively obviates the need for a comprehensive disclosure of all relevant medical information to the patient, the psychiatric consultant cannot establish and document that lack of capacity unless the treater has already provided that information.

Once the treating physician has disclosed the relevant medical information and documented the patient's treatment refusal, the psychiatrist can begin to assess the patient's epistemic and moral agency. Unfortunately, capacity assessments can be a philosophically convoluted process. Because there are no firm standards by which to ascertain when a patient has effectively acquired, retrieved, or meaningfully utilized medical knowledge, it remains indeterminate what must be in place for a patient to have full, or even adequate, epistemic agency.

There are several tools available for use in capacity evaluations. The most commonly used is MacArthur Competence Assessment Tool for Treatment

(MacCAT-T) (Grisso and Appelbaum 2001). This instrument identifies and utilizes four abilities central to medical decision making: (1) understanding medical information disclosed by the treating physician, (2) appreciating the relevance of that information to one's own situation, (3) reasoning about the decision in a manner that observers would consider rational and meaningful in light of one's personal values and past life decisions, and (4) making and communicating a choice. In considering the five elements of informed consent earlier, we saw that disclosure and capacity are preconditions, and now the MacCAT-T criteria make the "understanding and reasoning" component part of capacity. For this reason, some authors no longer include the "reasoning and understanding" element among the five components of informed consent.

The four abilities identified in the MacCAT-T resonate well with our view of epistemic and moral agency. Understanding medical information, appreciating its relevance to one's own situation, reasoning about how treatment choices would affect one's self and one's life, and authorizing or refusing treatment voluntarily and according to these processes and one's own values are exactly what moral and epistemic agents do. The psychiatrist's task, then, is to establish whether the person refusing a treatment recommendation can comprehend and utilize information to make her own decisions according to her own values and commitments.

Therefore, if the treating physician has adequately educated the patient about her illness and the available treatment options, thus ensuring that there is no coercion brought about by restricting medical information, the job of the consulting psychiatrist is to assess epistemic agency. The psychiatric interview is used to ascertain what the patient recalls and understands about her illness and treatment options; what that means in light of her own medical situation, life circumstances, and values; whether she can reason about those considerations in a manner someone else can follow and understand; and whether she can make a choice about what to do. Sometimes this process will culminate in a complete refusal of treatment, or a refusal until another time. Sometimes this process will culminate in treatment authorization. Sometimes the psychiatrist will find that there is a failure of agency, possibly because the physician has not sufficiently disclosed the medical information necessary for the patient to be a full moral and epistemic agent, and possibly because the patient's degree of epistemic agency at the time of the psychiatric interview prevents her from applying that medical information to her own situation.

Capacity assessments can be a formidable responsibility. Fortunately, capacity for treatment decisions is not all-or-nothing. Refusing a treatment with high

risks in favor of a more conservative treatment alternative, such as refusing surgery but accepting a noninvasive medical treatment, does not require the patient to exhibit the *degree* of capacity that might be required of someone refusing a treatment for which there is no alternative, or that constitutes the standard of care. This convention has been called the "sliding scale" of capacity (Drane 1985) to capture the idea that as the potential consequences of a patient's treatment refusal become more serious (i.e., higher risk of disability or death), stricter standards for capacity apply. For instance, refusal of a low-risk procedure such as phlebotomy and routine lab testing would not require the same degree of reasoning ability as would the refusal of lab testing to confirm or refute a diagnosis that is potentially morbid or lethal, that would be expected to require further treatment, such as sepsis. Conversely, refusal of a procedure with a high risk of morbidity or mortality, such as a surgery, or a procedure without well-established benefit, would not require as high a threshold for decision making capacity as a low-risk procedure or one with a higher success rate. In other words, the psychiatric consultant needs to have higher standards for establishing capacity for treatment refusal when the potential threat to the patient's well-being is high and the treatment is proven to be effective. That said, some authors argue for a unitary standard of capacity (c.f. Drane 1985), and even among authors who support a sliding scale, there is little agreement about which standards of informational comprehension or reasoning ability suffice for different patients facing various treatment decisions.

The psychiatric consultant who finds that capacity for decision making about a procedure is compromised may recommend ways to correct the limitation. The psychiatrist may be in a position to remind the treating physician that correcting metabolic derangements, treating infections and pain, soothing anxiety, providing social support, and other corrections can constitute step-wise measures to restore capacity for making the treatment decision at hand. When the patient mistrusts the treating doctor, sometimes the psychiatrist can reassure a patient about why the doctor recommends the treatment under consideration, or help the patient understand medical terminology, although these are not meant to be the psychiatrist's responsibility. A more collaborative relationship can also be established by including family or friends in the decision, or perhaps a hospital ombudsperson, ethics consultant or committee, or chaplain. It is important to remember that informed consent is an educational process that takes place over time, but that capacity is determined with respect to a given procedure at a specific point in time. Re-assessments of capacity may yield different results.

The role of the consulting psychiatrist is never to convince the patient to agree to the proposed treatment. Sometimes psychiatrists are asked to consult on a case with the expectation that the patient will be found to lack capacity to refuse treatment, so that the treating physician may proceed as planned over the patient's objection. When the recommended treatment is meant to restore capacity, there may be times when this is morally permissible, but one cannot assume that permissibility in all cases. Sometimes the treating physician expects the psychiatrist to document lack of capacity to refuse treatment and to convince the patient not to protest the treatment. As mentioned above, such an agreement is an assent, which is always a failure of informed consent. Assent to treatment is never the consulting psychiatrist's goal or responsibility.

Informed consent in psychiatric treatment

In the previous section, we discussed the role of psychiatric consultants to ascertain capacity for informed consent in acute medical settings. In this section, we explore the philosophical and practical challenges to informed consent when psychiatrists are the care providers rather than the consultants.

We want to be very clear that mental illness and psychiatric symptoms do not necessarily or universally interfere with a person's epistemic or moral agency. Mental illnesses and psychiatric symptoms can interfere with both, although not all conditions create such impairments, and when they do, they may do so to varying degrees. Furthermore, degrees of impairment can change over time, both with and without psychiatric treatment. Many persons who struggle with psychiatric problems are fully capable of participating in the moral and epistemic communities, and those whose symptoms interfere with innate abilities to act volitionally, reason clearly, understand treatment circumstances, or abide personal and community values are often only temporarily or situationally impaired. Given the many permutations of type and severity of psychiatric problems, there is no reason to question a person's epistemic or moral agency based solely on a diagnosis or the presence of symptoms.

Hence, informed consent is as central to treatment in psychiatry as it is in other medical disciplines.

However, since psychiatric symptoms can impair the capacity to make reasoned decisions about treatment, psychiatric practice brings philosophical questions about informed consent to the fore. Furthermore, the same symptoms can make it difficult for patients to trust psychiatrists, which makes it that much

more difficult to establish the collaborative relationship in which true informed consent can take place. Common problems such as depression, mania, anxiety, paranoia, command hallucinations, sleep deprivation, substance intoxication and addiction, compulsions, malnutrition, and cognitive impairments can limit both the epistemic and moral agency necessary for informed consent at various points in psychiatric patients' illnesses. These limitations make informed consent for psychiatric treatment extremely challenging, since the point of psychiatric treatment is often to restore that agency.

With vacillating capacity for informed consent, the key to maximizing informed consent in psychiatry is to build a mutually trusting therapeutic alliance with the patient and work toward an ongoing educational collaboration. In psychiatry, as in other medical specialties, disclosure and capacity are essential preconditions for informed consent. However, capacity must be considered, if not formally evaluated, for almost every clinical decision in psychiatry.

Before discussing other elements of informed consent, however, it is essential to address the element of informed consent that is far more problematic for psychiatry than for other medical disciplines: voluntariness. Society coerces treatment for persons with mental illness whose symptoms make them dangerous to themselves or others, and like it or not, psychiatrists are the vehicles of that coercion. Despite legal reforms in the 1960s and 1970s that limit the scope of psychiatry's power for involuntary psychiatric hospitalization and treatment, psychiatry still holds the social role of protecting both mentally ill individuals and society at large from violence or neglect resulting from psychiatric symptoms. Consequently, some academic commentators wonder whether informed consent is even relevant to psychiatric practice. We believe that it *is* relevant, yet psychiatry has a long history of coercive treatment, and as we have seen, without voluntariness, informed consent is moot.

Psychiatrists strive to collaborate with patients as much as possible, but involuntary hospitalization automatically creates an adversarial relationship between doctor and patient. Involuntarily committed patients know that the psychiatrist, or at least the institution, has divided loyalties between the patient and the state. Making treatment decisions together during involuntary hospitalization does not solve the problem, since the situation itself is coercive. Medication given over a patient's objection, or emergency sedation used to control a patient's aggressive behavior is always coercive, whether or not it is medically justified. Even when involuntarily committed patients do choose to participate in treatment, the involuntary hospitalization precludes the possibility

of informed consent. The aim of involuntary treatment may be to deny voluntariness to restore full moral agency, but it precludes informed consent.

The justification for involuntary treatment is that persons whose symptoms are severe enough to make them dangerous lack the capacity to refuse treatment. Recall that with respect to informed consent, capacity for decision making is always determined in reference to a specific decision at a specific time. We saw that if a person is able to (1) understand her illness and her treatment options, (2) appreciate the relevance of that knowledge to her own situation, (3) reason about treatment in light of her own values and life goals, and (4) make a choice about treatment, then she has the capacity to authorize or refuse treatment. If that person's symptoms genuinely interfere with any of these elements, a fully informed consent from the patient is not required. Persons whose illnesses cause suicidality, self-injury, self-neglect, or violence toward others may not understand that these impulses are caused by potentially reversible psychiatric symptoms, since the phenomena often feel authentic and well-integrated with the rest of the person's life. Values and life goals may dramatically change with mood, eating, and psychotic disorders; understanding and self-awareness can alter with a variety of mental disorders; and insight, judgment, self-control, and self-directed choice can vary in any number of ways. The paradoxical assumption in psychiatry and in society generally is that a person who is dangerous to herself or others lacks capacity to refuse hospitalization, and that capacity should be restored to whatever extent possible, even if doing so restricts self-determination and moral agency temporarily. This assumption, and the practices that follow from it, do harm people, but American society has yet to respond with a better alternative to involuntary psychiatric hospitalization.

We presented informed consent early in this chapter as a collaborative process that takes place over time, an ongoing discussion during which different decisions can be made, reconsidered, and revised. This process requires trust between doctor and patient, and involuntary hospitalization can make it difficult for patients to trust psychiatrists. People who previously have experienced real or perceived harm during psychiatric treatment (e.g., people who sometimes call themselves "psychiatric survivors," "service users," "consumers," or "clients" to distance themselves from psychiatry's claim of authority) may experience treatment to be manipulative or coercive because it actually has been, or sometimes because psychiatric recommendations feel like external loci of control, or judgments about who patients are as people or how they live their lives. Psychiatrists, especially trainees and early career psychiatrists, may not recognize that despite efforts to earn and be worthy of patients' trust, patients

may view psychiatrists with suspicion. The first challenge to informed consent in psychiatry, then, is psychiatry's own practice of involuntary hospitalization and treatment.

In less acute, voluntary treatment settings—e.g., outpatient clinics, day programs, substance rehabilitation programs—it is easier to establish and earn patients' trust. Psychiatrists must still face the usual challenges of patient defenses (i.e., externalization, resistance, denial, projection, projective identification) and unconscious wishes and fantasies, but when voluntariness is restored, and moral agency respected, psychiatrists can begin to establish the trusting, collaborative relationships that informed consent requires. Successful voluntary psychiatric treatment respects patients' moral agency, and works toward maximizing epistemic agency in a way that is not always possible for patients who are acutely symptomatic.

Psychiatrists face the second challenge to informed consent when they undertake disclosure obligations: most psychiatric treatments are not benign, and some have serious, lasting side effects that must be disclosed at the beginning of treatment. Many psychiatric treatment options have potentially dangerous or unpleasant short- and long-term effects that must be revealed to patients as part of the informed consent process. The paradox and the frustration of informed consent in psychiatry, as in any other medical specialty, is that full and proper disclosure of potential risks and benefits of a given treatment can result in an informed refusal of treatment rather than consent.

Finally, the third challenge to informed consent in psychiatry is that certain psychiatric symptoms can make the disclosure element of informed consent more difficult. Changes in mood, perception, self-worth, insight, decisiveness, and others can all influence a patient's willingness to participate in treatment, even if they do not compromise capacity. Anxiety may prevent someone from taking anxiolytic medication; depression may make someone hopeless that any treatment attempt is worthwhile; auditory hallucinations may provide contrary information about an antipsychotic medication; the potential for overdose may make a suicidal patient refuse one treatment for a potentially toxic other; fears about weight gain can keep someone with anorexia from participating in treatment at all; paranoid delusions may find purchase on the risk of dystonic side effects or metabolic syndrome, etc. But short-term and long-term adverse effects of treatments do commonly occur, and their possibilities must be conveyed to the patient. The art to disclosure in psychiatry is to be truthful without minimizing treatment risks for the anxious or overstating treatment benefits for the hopeless, and to take seriously a

patient's fears (whether or not they are based in psychopathology) without catering to them completely.

Again, the solution lies in the therapeutic relationship, and its development over time. Psychiatric treatment decisions are seldom like decisions to undergo surgery: it is not usually the case that a single authorization is made and by the time the patient reconsiders, the deed is done. Most pharmacologic and psychotherapeutic treatment decisions can be reevaluated, long-term adverse effects emerge over weeks to years, and ambivalent patients often appreciate reassurance that they may change their minds for any reason at any time. Voluntary hospitalization authorizations may be rescinded, although most states allow a few days delay to assess for safety and allow time for aftercare planning. Electroconvulsive therapy must be authorized for each treatment. As a patient comes to trust himself as a decision maker, and to trust the treatment provider as one who respects his agency as a knower and a decision maker, decisions can become more collaborative.

In summary, informed consent for treatment in psychiatry can be more challenging than it is in other disciplines because involuntary treatment erodes trust in the profession. Furthermore, many psychiatric treatments have harmful sequelae, and psychiatric symptoms can make decision making difficult for patients. As frustrating as this may be, when trust is established, the investment in the therapeutic relationship makes informed consent possible and can serve as its own reward.

Informed consent for psychiatric research

It is important to grasp the differences between informed consent in clinical and research settings. Whereas psychiatric treatment is intended to directly benefit the patient by controlling uncomfortable or frightening symptoms or improving general role functioning, psychiatric research is not performed for the primary purpose of benefiting an individual patient, but has as its overriding goal the development of generalizable knowledge. Informed consent in research conveys respect for individual subjects, as it serves to notify people that research is ongoing and that they have the right not to take part. Informed consent also promotes subject safety, by alerting subjects to risks and uncertain outcomes. Perhaps most importantly, informed consent is an express limitation on investigators' authority, as researchers must approach potential subjects with hat in hand, prepared to convince each person that taking part will not be overly risky or burdensome, can help understand or alleviate a problem or contribute

meaningfully to science, and is a reasonable thing to do. Because as a general rule people cannot be compelled or hoodwinked into research participation, researchers must afford each potential subject the ability and freedom to refuse participation, for any reason, at any time.

The legal and ethical roots of informed consent for research date from at least the end of the nineteenth century. Public outcry followed revelations of syphilis experiments performed on prostitutes and others by Albert Neisser in Breslau, Germany, which led in 1900 to Prussian regulations requiring "unambiguous consent" from competent subjects to any "medical interventions for other than diagnostic, healing, and immunization purposes" (Vollmann and Winau 1996a). Around the same time, experimentation on transmission of Yellow Fever was performed in Cuba by Walter Reed and colleagues. A consent form was used which made clear that there was a risk of death, offering $100 in gold for participation and another $100 in case the subject contracted Yellow Fever (payable to relatives if the subject succumbed to the disease) (U.S. Army Yellow Fever Commission 2017).

Revelations about problems spurred later legal changes. In Germany, public concerns were sparked by distribution of a contaminated tuberculosis vaccine that killed some 29 percent of vaccinated children. This led to adoption of broad regulations of research in Germany, specifically requiring informed consent (Bonah 2002). These were the first national regulations dictating the ethics of research on humans, and it was ironic that it was German doctors who performed gross inhuman research on prison camp victims during the Second World War (Vollmann and Winau 1996b). The war crimes trial of some of those doctors at Nuremberg was not focused on their violation of the German law, nor on their crimes against humanity. Instead, the judges included in their final decision a "code" of ten rules that they believed provided the standards for performing research on human subjects. The first and universally followed rule requires informed consent from subjects (Shuster 1997).

International guidelines have been adopted in the decades since, articulating a clear standard for securing consent from subjects for research. The Declaration of Helsinki is guidance provided by the World Medical Association. First promulgated in 1964, the Declaration has been revised more than a dozen times since (World Medical Association 2013). More recently, the Council for International Organizations of Medical Sciences promulgated its own set of guidelines, targeted to countries that do not have the resources to develop their own regulations. This guidance was updated in 2016 (Council for International Organizations of Medical Sciences 2016).

In the United States, the Nuremberg Code, issued by a duly convened military tribunal, was binding upon only the military. Thus, there was little law in the United States that would require physicians (or other researchers) to secure the informed consent from their subjects. In 1953, with the opening of the new National Institutes of Health clinical center, a Medical Board policy was adopted that required, among other protections, informed consent from patient-subjects (NIH 1953). This policy was expanded in early 1966 to require review of proposed research by a committee of institutional associates (later termed an Institutional Review Board (IRB) in the United States, or Research Ethics Boards or Committees in other countries) and informed consent from subjects for both intramural NIH research and research by Public Health Service grantees (U.S. Public Health Service 1966). The NIH issued guidance for protection of human subjects in 1971 (U.S. Department of Health, Education, and Welfare 1971). Then, following revelations about the study of untreated syphilis in African American citizens in Alabama (the so-called Tuskegee Syphilis Study), Congress enacted the National Research Act in 1974 (National Research Act 1974). Simultaneously, the Department of Health, Education, and Welfare issued regulations essentially codifying the 1971 guidance document.

The Act also established the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The National Commission, as it is commonly referred to, met numerous times over the following five years, commissioned a large number of papers and studies, and generated a thoughtful and highly influential final report, The Belmont Report (National Commission 1979). This concise report is perhaps the single most important document that a budding biomedical researcher should read.

The regulations have been amended numerous times over the years. In 1991, the rules were formally adopted by all federal agencies that fund or perform human subjects research, and since then the basic rule (subpart A) has been referred to as the Common Rule (DHHS 2018). This Rule was given its first major overhaul in early 2017 (DHHS 2017). Other parts of the regulation apply to research on pregnant women, prisoners, and children. Notably, no special rules were issued for research on persons with mental illness or cognitive limitations, though researchers and institutional review boards consider such subjects to be vulnerable and deserving of special protections. What form such protections take and how rigorously the research enterprise pursues these obligations are largely unknown, but recent abuses that have been brought to light suggest there are problems (Elliott 2010).

Guidelines for informed consent in research uniformly demand much more comprehensive disclosures than in any clinical setting. As discussed above, research typically asks subjects to perform tasks and accept risks not for their own benefit, but for the benefit of others or to contribute to scientific knowledge. Thus, it is uniformly agreed that potential subjects must be made aware of the known, foreseeable risks of participation, including the fact that there may be risks that are simply unforeseen, given the nascent state of the science or limited experience with the subject matter under study. Informed consent must be secured in a way that maximizes free choice by potential subjects, avoiding any coercive threats that would leave subjects worse off if they were to refuse, and avoiding the offer of any undue incentives (e.g., housing, excessive payments, or "free treatment") to entice subjects.

Psychiatric research presents various challenges stemming from the affective, psychotic, cognitive, attentional, and other symptoms experienced by potential subjects. Whether subjects have capacity to consent to research is a core question researchers must face. How carefully must capacity be assessed to enable a potential subject to join a study, and by what standards? By whom should capacity be assessed? Should a researcher, who likely is highly motivated to enroll subjects, be entrusted with this? If a potential subject is determined to lack capacity, who—if anyone—may give permission on behalf of a patient? What is the standard a legally authorized representative should apply when deciding, a substituted judgment (what would the patient want if she were able to decide now?) or best interest (what would be in the overall best interest of the patient?)?

A related problem arises once a subject is enrolled. Subjects have an unfettered right to withdraw from research at any time, for any reason. Persons experiencing psychiatric symptoms may be unable to assess their continued willingness to take part, and indeed, to even monitor how they are being treated by the research team. This may be compounded for trials that involve changing a subject's treatment, including in particular trials that involve a wash-out period, from which time subjects may suffer relapses that diminish that person's ability to defend her rights and welfare if research subject protections are not respected.

Conclusions

Informed consent is a challenging and sometimes controversial topic in any field of biomedicine, but psychiatry adds unique challenges. Informed consent is an

ongoing collaborative procedure, grounded in a trusting therapeutic relationship, that allows voluntary authorization of medical treatment in light of one's own illness, need for treatment, treatment options, and the risks and potential benefits of each. We have described this capacity for self-determination in terms of moral and epistemic agency, where agency is simply the ability to act on one's own behalf. Recognition of moral agency acknowledges persons as members of a shared moral community; recognition of epistemic agency acknowledges persons as members of a shared community of knowers.

The law does not provide real direction for physicians. Not to obtain any consent constitutes a battery, but the law typically considers insufficient informed consent a matter of malpractice. Some states have undermined the ethical aim of informed consent—to foster better communication between physicians and patients rather than to just document a disclosure conversation—rather than supported it. We encourage doctors to learn their states' informed consent requirements, but to uphold the more demanding ethical standards, pursuing good doctor-patient two-way communications to avoid misunderstandings and false expectations, and try to prevent such suits in the first place.

We addressed capacity as one of the challenges facing psychosomatic psychiatrists. Despite that the definition of capacity is circular, that there are no clear standards for determining it, and that consulting physicians often don't fulfill their obligations of disclosure before requesting capacity assessments, psychosomatic psychiatrists regularly make capacity determinations. Capacity is a precondition for valid informed consent, in the sense that a patient must have full epistemic agency to voluntarily authorize or refuse treatment. In ascertaining capacity for making informed treatment decisions, psychiatrists often must identify situational and reversible factors that temporarily impair epistemic agency, and must strive to understand when a patient merely assents to treatment rather than giving a full informed consent. In cases of treatment refusal, the psychiatrist should keep in mind that there may be multiple reasons for treatment refusal and that lack of epistemic agency—understanding or reasoning—may not be the primary problem if there are communications problems between the patient and the primary care team that the patient perceives to limit moral agency and self-determination.

Informed consent is as central to treatment in psychiatry as it is in other medical disciplines. Unlike other disciplines, however, the history of involuntary treatment in psychiatry casts a long shadow over the trusting doctor-patient relationships necessary for informed consent. When treatment is involuntary, informed consent is not possible, since voluntariness is one of the five constitutive

elements of informed consent and part of moral agency. When treatment is voluntary, people with histories of involuntary treatment may justifiably experience lingering mistrust. Trust and education may also be hard won for patients whose psychiatric symptoms made them wary, ambivalent, indecisive, paranoid, cognitively impaired, or limited by other symptoms that treatment is meant to address. In these cases, patience and ongoing efforts to restore and affirm epistemic and moral agency over time can be rewarding. Reminding the patient of his own ability to ask questions, halt or defer treatment, or change his mind often provide reassurance about the relationship and his own agency that make informed consent for psychiatric treatment possible.

In psychiatric research settings, informed consent is used to protect research subjects by notifying them of predictable risks and potential benefits of research participation, limiting the authority of the investigators, and assuring subjects that they may withdraw from participating at any time, for any reason. Accordingly, disclosures for informed consent in research must be much more comprehensive than those for medical treatment, which is intended to benefit the patient, whereas research is aimed primarily at generating new knowledge. Capacity to consent to research participation must be assessed more carefully, since a subject should be symptomatic enough to quality for participation in the study, yet not so symptomatic as to lack capacity to give informed consent for that participation. Depending on the study design, subjects may become more symptomatic and lose capacity for continued participation during the research protocol, which raises questions about ongoing assessment of capacity throughout the course of the study and responsibilities of others for protecting subjects unable to protect themselves.

Acknowledgments

The authors thank Michele Daniele, Esq. and Lea Hecht, M.D. for helpful insights and direction.

Note

1 The Supreme Court of Washington held in *Helling* that reasonably prudent opthalmologists must perform intraocular pressure testing on patients younger than forty years of age, determining that, as a matter of law, the potential benefits of

avoiding blindness from undetected but extremely unlikely glaucoma outweighed the minimal cost of testing. The Court's decision was highly controversial, taking much too simple a view of diagnostic uncertainty, and its willingness to conclusively determine the legal adequacy of medical practices has not been followed by other courts (Fortess and Kapp 1985: 212).

References

Appelbaum, P. S. and Gutheil, T. G. (2007), *Clinical Handbook of Psychiatry and the Law*, 4th edition. Philadelphia, PA: Wolters Kluwer.

Beauchamp, T. L. and Childress, J. F. (2001), *Principles of Biomedical Ethics*, 5th edition. New York: Oxford University Press.

Bonah, C. (2002), "Experimental Rage": The Development of Medical Ethics and the Genesis of Scientific Facts, *Social History of Medicine*, 15:187–207.

Canterbury v. Spence, 464 F.2d 772, 783 n. 36 (D.C. Cir.), cert. denied, 409 U.S. 1064 (1972).

Cobbs v. Grant, 8 Cal.3d 229, 104 Cal. Rptr. 505, 502 P.2d 1 (1972).

Department of Health and Human Services (2017), Federal Policy for the Protection of Human Subjects, *Federal Register*, 82: 7149–7274, January 19, 2017.

Department of Health and Human Services (2018), Protection of Human Subjects. 45 CFR Part 46.

Drane, J. F. (1985), The Many Faces of Competency, *Hastings Center Report*, 15:17–21.

Elliott, C. (2010), The Deadly Corruption of Clinical Trials, *Mother Jones*, September/October. URL= http://www.motherjones.com/environment/2010/09/dan-markingson-drug-trial-astrazeneca/ Retrieved July 23, 2017.

Faden, R. R. and Beauchamp, T. L. (1986), *A History and Theory of Informed Consent*, New York: Oxford University Press.

Fortess, E. E. and Kapp, M. B. (1985), Medical Uncertainty, Diagnostic Testing, and Legal Liability, *Law Medicine & Health Care*, 13:213–218.

Grisso, T. and Appelbaum, P. S. (2001), *MacArthur Competence Assessment Tool for Treatment (MacCAT-T)*, Sarasota, FL: Professional Resource Press.

Helling v. Carey, 83 Wash. 2d 514, 519 P.2d 981 (1974).

Hobbes, T. ([1668] 1994), *Leviathan*. E. Curley, ed. Indianapolis, IN: Hackett Publishing Company, Inc.

International Ethical Guidelines for Health-Related Research Involving Humans (2016), 4th edition. Geneva: Council for International Organizations of Medical Sciences (CIOMS).

Kant, I. ([1785] 1993), *Grounding for the Metaphysics of Morals*, 3rd edition. J. W. Ellington, trans. Indianapolis, IN: Hackett Publishing Company, Inc.

Karlawish, J. H. (2016), *Assessment of Decision-Making Capacity in Adults*, Last modified September 9, 2016. Retrieved July 24, 2017.

Katz, J. (1984), *The Silent World of Doctor and Patient*, New York: The Free Press.

Kutner, M., Greenberg, E., Jin, Y., and Paulsen, C. (2006), *The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy* (NCES 2006-483), Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Meisel, A. and Kabnick, L. D. (1980), Informed Consent to Medical Treatment: An Analysis of Recent Legislation, *University of Pittsburgh Law Review*, 41:407.

Meisel, A. and Kuczewski, M. (1996), Legal and Ethical Myths about Informed Consent, *Archives of Internal Medicine*, 156:2521–2526.

Merz, J. F. (1991), An Empirical Analysis of the Medical Informed Consent Doctrine: The Search for a "Standard" of Disclosure, *Risk: Health, Safety & Environment*, 2:27–76.

Merz, J. F., Fischhoff, B., Mazur, D. J., and Fischbeck, P. S. (1993), A Decision-Analytic Approach to Developing Standards of Disclosure in Medical Informed Consent, *Journal of Products and Toxics Liability*, 15:191–215.

Natanson v. Kline, 186 Kan. 393, 350 P.2d 1093 (1960), modified, 187 Kan. 186, 354 P.2d 670 (1961).

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979), *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*, Washington DC, Department of Health, Education, and Welfare. URL= https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/ Retrieved January 27, 2017.

National Institutes of Health (1953), *Medical Board Document No. 1: Organization, Functions and Authority of the Medical Board of the Clinical Center*. National Research Act, Pub. L, 93–348, July 12, 1974, Bethesda, MD.

Salgo v. LeLand Stanford Jr. Board of Trustees, 154 Cal. App. 2d 560, 317 P.2d 170 (1957).

Schloendorff v. The Society of the New York Hospital, 211 N.Y. 125, 105 N.E. 92 (1914).

Shuster, E. (1997), Fifty Years Later: The Significance of the Nuremberg Code, *New England Journal of Medicine*, 337:1436–1440.

Tauber, A. I. (2005), *Patient Autonomy and the Ethics of Responsibility*, Cambridge, MA: The MIT Press.

U.S. Army Yellow Fever Commission. URL= http://exhibits.hsl.virginia.edu/yellowfever/ Retrieved January 26, 2017.

U.S. Department of Health, Education, and Welfare, Public Health Service, National Institutes of Health (1971), *The Institutional Guide to DHEW Policy on Protection of Human Subjects*, Washington, DC: DHEW Publ. No. (NIH), 72–102.

U.S. Public Health Service, Division of Research Grants. Policy PPO #129, February 8, 1966.

Vollmann, J. and Winau, R. (1996a), The Prussian Regulation of 1900: Early Ethical Standards for Human Experimentation in Germany, *IRB*, 18:9–11.
Vollmann, J. and Winau, R. (1996b), Nuremberg Doctors' Trial: Informed Consent in Human Experimentation before the Nuremberg Code, *BMJ*, 313:1445–1447.
World Medical Association (2013), *Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects*. URL= http://www.wma.net/en/30publications/10policies/b3/ Retrieved January 26, 2017.

14

The Ethics of Coercion and Other Forms of Influence

Kelso Cratsley

Introduction

Across the health sector there is increased recognition of the ethical significance of interventions that constrain or coerce. Much of the recent interest stems from debates in public health over the use of quarantines and active monitoring in response to epidemics, as well as the manipulation of information in the service of health promotion (or "nudges"). But perhaps the area in which these issues remain most pressing is mental health, where the specter of involuntary treatment has always loomed large. Indeed, there are good reasons to think of psychiatry as ethically exceptional given its broader role as social and legal arbiter (Radden 2015). This is compounded by the fact that mental health treatment has other special features that can serve as potential entry points for more subtle forms of pressure and influence. These include the relative cognitive and emotional vulnerability of patients, the highly confidential information involved, and the importance of therapeutic relationships. Thus the field of mental health needs to grapple with a range of closely related issues concerning the ethics of influence.

In this chapter I explore several, with a primary focus on clinical concerns, although a number of points have relevance for research ethics. Discussion of treatment pressure has traditionally centered on *coercion*, or, very roughly, the use of threats or force to get someone to do something they otherwise would not. Here I follow suit, beginning with examination of its conceptual standing alongside other categories of influence, including *manipulation* and *persuasion*. As it stands, there is nothing close to a consensus on these matters (reviewed in Anderson 2011), but a relatively broad understanding of coercion can still be useful

for ethical deliberation. There is also a growing body of empirical literature on coercion in psychiatry, including research on the use of coercive measures such as involuntary hospitalization, physical restraint, and forced medication, as well as the construct of "perceived coercion," which tracks patients' experiences of treatment pressure (reviewed in Kallert et al. 2011a; Molodynski et al. 2014; Newton-Howes and Mullen 2011). I briefly describe several key findings and their relation to specific recommendations to reduce coercion.

Building on all of this, I then take up the normative questions directly by evaluating a recently proposed framework for the ethical analysis of influence (Blumenthal-Barby et al. 2013). I recommend that mental health continue to focus on the question of coercion and related issues concerning autonomy, capacity, and consent familiar from debates over involuntary treatment. However, I also suggest that recent findings underscore the importance of more subtle forms of influence. At the core of these matters are questions about the nature of relationships—between patients, practitioners, family members, and even with broader social norms—and the ways in which information is framed and processed. At the very least, the ethics of persuasion are far from straightforward given disparities in authority and capacity. As in current debates in public health, mental health may also need to attend to the ways in which the manipulation of information can effectively bypass rational assessment.

Theoretical background

Any discussion of these issues must start with the concept of coercion, given its long and contentious history in political philosophy, legal theory, and biomedical ethics. The aim here is not to attempt a full treatment, merely to describe its relevance for mental health ethics. A broad working definition is necessary, though, if only to mark it out from closely related notions such as compulsion, deception, and inducement. There is another important motivation. All too often the term coercion is applied to certain types of intervention as a signal of ethical disapproval, presumably in response to constraints on autonomy or freedom of choice but absent a specific definition or detailed argument. This problem has plagued discussions of nudges in public health (Blumenthal-Barby 2012; Quigley 2014), and something similar regularly occurs in research ethics (Hawkins and Emanuel 2005; Largent et al. 2013). Debates in mental health are not immune to this problem, so a relatively clear conception is required (Szmukler and Appelbaum 2008; Szmukler 2015).

As it happens, leading theorists of coercion have expressed distaste for definitions while at the same time offering their preferred formulation. The most influential account is no exception—Nozick (1969) famously presented a list of necessary and sufficient conditions qualified with caveats about its limitations. In simplest terms, Nozick defines coercion as the intentional use of a credible threat to influence or control the actions of another person.[1] This definition has proven remarkably durable and influential, as it captures a fairly intuitive notion of coercive threat. Take the example of someone holding a gun to your head and stating "Your money or your life." Highlighting some of Nozick's key criteria, there is a *conditional threat*, the threat is *credible*, the coercee makes a *choice* of sorts whether to acquiesce, and resistance to the threat would leave the coercee *less well off*. This last criterion is particularly important because many definitions appeal to a prior baseline to which the outcome of coercive threats can be compared. That is, instances of coercion harm the coercee in some significant way. This can also help distinguish coercive threats from non-coercive offers, as the latter presumably have little impact if the coercee turns them down (even if problematic in some other way).

Other theories of coercion have also relied upon the notion of a baseline. Most notably, Wertheimer (1987) required that, first, the threatened outcome must leave the coercee morally worse off, and, second, the coercee has no real choice but to submit and is therefore not responsible for their actions. This approach thus places a premium on the possibility of mitigated responsibility. Standard examples include inadmissible confessions, where information acquired through forced interrogation can be thrown out of court, and breaches of contract, where contracts signed under duress can be voided. It also means that offers, unlike threats, are rarely, if ever, coercive. Even in instances where declining an offer leaves the decliner less well off morally, they are almost always responsible for their decision. Consider plea bargains, offers that may leave little room for refusal but not in a way that invalidates the content of the plea itself. Thus in research ethics Wertheimer maintained that exploitative offers to participate in research are never "coercion" but better understood as "undue inducement" (Wertheimer and Miller 2008).[2]

For the purposes of mental health, these theories have their advantages. Wertheimer's account foregrounds the status of consent under threat, thereby prioritizing what is arguably the most ethically significant question. Evaluations of coercive measures in mental health, then, should focus on the "morally relevant distinctions" between interventions that do or do not undercut consent (Wertheimer 1993). While involuntary commitment may remain paradigmatic

of coercion, ethical analysis will be best served by attending to debates over what—if anything—distinguishes it from, say, familial pressure, which itself may constrain patient choice in morally significant ways. For example, if someone with a substance misuse disorder is pressured by their family into entering treatment, they may very well feel that they were in some ways forced into treatment, while others might maintain that the patient was still ultimately responsible for their decision (and thus not "coerced"). In addition, both Wertheimer and Nozick frame coercion as a matter of interpersonal threat. This jibes well with some of the special features of mental health, including the importance of the clinician-patient relationship. However, there are several limitations to this general approach, largely having to do with the range of pressures considered. In short, it may be insufficiently inclusive.

Varieties of influence

For a start, a number of commentators have held that offers can in fact be coercive. Perhaps most influentially, Feinberg (1986) has argued that certain offers undermine consent if made in circumstances fashioned for the very purpose of taking advantage of the recipient. For example, imagine the case of a prison warden who offers commutation of a death row inmate's sentence in exchange for participation in a medical experiment. The broader point is that certain arrangements may constrain an individual's range of choices to such an extent that consent is rendered invalid, particularly if such circumstances include a significant power differential. This clearly has relevance for research ethics, where others have made similar arguments, *pace* Wertheimer (McMillan 2010).[3]

Another limitation of the threat-based account is that it neglects the more traditional view of coercion as the use of physical force, familiar from the work of political philosophers who equated coercion with state-sanctioned impositions of force (e.g., Hobbes, Kant, and Mill). The concern is that a purely threat-based account, or the "pressure" view, misses important aspects of coercion related to the application of force, better captured by the "enforcement" view (Anderson 2010). Most importantly, the latter emphasizes the background conditions that allow for coercion in the first place, such as the coercer's position of power or possession of the means for following through on threats. Think back to the example of armed robbery; or, more generally, incarceration; or, more topically, "enhanced interrogation." And so while the threat-based view typically assumes that coercion acts on the *voluntariness* of the coercee, altering their will through

constrained choice, the enforcement view stresses that being in a position to carry out threatened actions is essential.

These points have obvious application to mental health, where interventions such as involuntary hospitalization or physical restraint are forced on patients who are in a relatively weak or vulnerable position. And if these points hold, then they expand the range of possible applications of the notion of coercion. At the very least, we shouldn't be too bothered with policing conceptual boundaries. But regardless of one's preferred conceptualization, it is necessary to recognize the diverse forms of influence at work in health care, including mental health. For instance, a highly influential framework distinguishes the three broad categories of *coercion, manipulation*, and *persuasion*, which describe a continuum of controlling influence (Faden and Beauchamp 1986). Manipulation lies in the middle, covering deception, offers and inducements, the management of information, social and cultural norms, and interpersonal leverage. At the extremes are coercion, which is completely controlling, and rational persuasion, which is completely *non*controlling. Complete control occurs when someone is "entirely dominated by the will of another … through irresistible influences" (269). Coercion thus requires intent to coerce, a credible threat, and irresistible influence.

This framework provides a helpful way of thinking about various forms of influence. One clear implication is that coercion can nullify informed consent, but it also leaves open a wide range of influences that do not—i.e., the "territory of influences compatible with substantially autonomous acts" includes persuasion and substantially noncontrolling forms of manipulation (339). This approach is also intended to be normatively neutral. That is, locating an instance of influence within the typology is distinct from debating whether its use is justified. But of course it's not difficult to see how it directly relates to ethical concerns. In the context of mental health, a somewhat similar framework has been proposed, comprising a hierarchy of treatment pressures (Szmukler and Appelbaum 2008). These include: persuasion, leverage, inducements, threats (or coercion), and compulsion (or forced inpatient or outpatient treatment). Here the connection to normative claims is intrinsic, with the justificatory burden increasing with the degree of restrictiveness. This framework doesn't offer detailed ethical guidance, but it does provide another useful tool with which to approach the pertinent issues.[4] There are no easy answers to the normative questions in mental health, but there has been a recent upsurge of interest, much of it empirical in nature. I next turn to this research before taking up ethical guidelines in the final section.

Empirical evidence

There is now a large empirical literature on coercion in psychiatry, most of it originating in several international research initiatives (reviewed in Kallert et al. 2011a; Molodynski et al. 2014; Newton-Howes and Mullen 2011). Of special note is the EUNOMIA project, funded by the European Union across twelve different countries (Kallert et al. 2005), as well as the Nordic Coercion Study and Swedish Coercion Study (Hoyer et al. 2002; Kjellin and Wallsten 2010). Much of this research has been focused on establishing the prevalence, correlates, and associated clinical outcomes of coercive measures, such as involuntary commitment, seclusion, restraint, and forced medication. There is also ongoing work in the UK, with research groups focusing on qualitative research (Priebe et al. 2010a) and community treatment orders (Burns et al. 2013). And an important early forerunner was the MacArthur Foundation Research Network on Mental Health and the Law in the United States (Gardner et al. 1993).

This list of research programs is far from exhaustive, and what follows is a highly selective review of several key findings. As such, the aim of this section is to impart a basic sense of what the extant data can tell us about coercion and related forms of influence. The descriptive picture remains incomplete, and it is far from clear what implications it has for both our understanding of coercion and the justification of its use. It does, however, represent an important body of evidence from which to draw when evaluating conceptual and normative claims. These empirical findings may not map neatly onto traditional philosophical categories or commitments.[5] But they are interesting findings nonetheless, some of undeniable ethical significance. At minimum, they can help direct our moral attention.

Coercive measures

The prevalence of legally mandated inpatient treatment varies across countries and jurisdictions, unsurprising given the diversity of health systems and legal standards. Some countries have seen an increase in the use of involuntary inpatient treatment (de Stefano and Ducci 2008). For example, in the UK the compulsory admission rate has risen as the number of psychiatric hospital beds has decreased (Keown et al. 2011), while more than half of Mental Health Act evaluations result in legal commitment, associated with the following patient characteristics: severity of symptoms, psychotic disorder, and low global functioning (Singh et al. 2014). Once involuntarily hospitalized, roughly 40

percent of patients are subjected to coercive measures, though again there is wide variation (Raboch et al. 2010; Kalisova et al. 2014). In terms of the use of specific measures, there are few consistent patterns. For example, a Norwegian study (Husum et al. 2010) found seclusion to be the most common (35 percent), followed by restraint (10 percent), and forced medication (9 percent); whereas a EUNOMIA study (McLaughlin et al., 2016) found forced medication to be the most common (41 percent).[6] The use of these measures has been consistently linked to patient aggression, severity of illness, psychotic disorder, and low global functioning (Kalisova et al. 2014).

The standard justification for these interventions is that they are in the best interest of the individual patient and general public. This has given rise to an important strand of research focused on clinical outcomes of involuntary treatment. Some of this work has shown that a majority of patients ultimately show symptomatic improvement and a significant minority report satisfaction with their treatment. For example, an initial review found that inpatient commitment was associated with clinical improvement, including a range of patients (33–81 percent) who retrospectively described their treatment as justified or beneficial (Katsakou and Priebe 2006). One study found that only 15 percent of patients had been readmitted to hospital one year after their initial commitment and 40 percent considered their original admission to be justified (Priebe et al. 2009). Research within the EUNOMIA project reported that anywhere from 39 to 71 percent of patients described their inpatient commitment as justified after 1 month, increasing to 39–71 percent after three months (Priebe et al. 2010a), and significant symptomatic improvement after both 1 and 3 months (Kallert et al. 2011b).

This research has started to provide a clearer picture of coercion in mental health, and it may be somewhat reassuring that what is done in the name of better outcomes does appear to serve that purpose (at least for *inpatient* treatment).[7] But of course data alone cannot settle normative disputes. Several points are worth underlining. First, regardless of outcomes, one could still argue that certain types of intervention are ethically problematic—or even impermissible—no matter what. The substantial degree of infringement on autonomous choice might mean that some coercive measures are moral nonstarters. Second, regarding the retrospective views of patients after the initiation of treatment, it goes without saying that "retroactive approval" or gratitude is not consent (Wertheimer 1993). The findings here are instructive but not overwhelming, with a significant percentage of patients presumably continuing to *disagree* with the grounds for their detention. The fact remains that consent requires

an active, authorized decision prior to the initiation of treatment. This kind of temporal consideration highlights the appeal of psychiatric advanced directives or advanced statements (Appelbaum 2004). Lastly, a large proportion of involuntary patients are mentally capacitous at point of admission (Cairns et al. 2005). Therefore one could argue that such patients should retain the right to refuse treatment. Considerations of this kind are consistent with the suggestion that legislation governing involuntary treatment should incorporate standards regarding capacity (Szmukler 2015).

A less tendentious conclusion to draw from the data on coercive measures is simply that their utilization is reducible. The EUNOMIA project has produced recommendations to this effect. This starts with improved treatment of symptoms prior to hospitalization, given the robust association between severity of symptomatology and legal detention (Fiorillo et al. 2012). Other steps can also be taken, such as giving patients better information about their hospitalization, including its justification and duration; an increased focus on protecting patient rights; more family involvement; clearer lines of communication between the hospital and community; more meetings with service users; and better training of staff about interventions for aggressive behavior (Fiorillo et al. 2011). This last point is particularly important given data suggesting that ward staff don't have an adequate understanding of coercion (Jaeger et al. 2014), and evidence that staff training can improve the experiences of involuntary patients (Wykes et al. 2017). A recent systematic review found advance statements to be the only effective means of reducing involuntary commitment (de Jong et al. 2016).

Perceived coercion

The previous findings concern what is sometimes referred to as "objective" coercion, but there has also been a great deal of research on the subjective experiences of psychiatric patients. Most of this work has used the construct of "perceived coercion," first operationalized in a scale developed in research funded by the MacArthur Foundation (Gardner et al. 1993). This was driven by early data revealing that legal status was a poor gauge of coercion, as some *voluntary* patients reported coercive experiences while some *involuntary* patients denied having been coerced (e.g., Rogers 1993). Thus an alternative measure was deemed necessary, with perceived coercion defined as the "belief that one lacks choice, control and influence over admission to hospital, and doesn't have freedom to refuse" (Gardner et al. 1993). Researchers have also used the Nordic

Admission Interview, which centers on questions regarding patient perspectives on self-determination, awareness of legal status, restrictions on freedom, and violations of integrity (Hoyer et al. 2002).

Research using these measures has revealed that perceived coercion is quite common in involuntarily committed patients, with anywhere from 74 to 98 percent of patients reporting coercive experiences (Kjellin et al. 2006; O'Donoghue et al. 2014). A significant minority of voluntary patients report similar experiences, ranging from 10 to 34 percent (Bindman et al. 2005; Hoge et al. 1997; Iversen et al. 2002; Kallert et al. 2011b; Kjellin et al. 2006; O'Donoghue et al. 2014). In terms of patient characteristics, perceived coercion in both voluntary and involuntary patients tends to be associated with the positive symptoms of psychosis and poor global functioning (Fiorillo et al. 2012; O'Donoghue et al. 2014), while the data regarding age, gender, and ethnicity are decidedly mixed (Newton-Howes, and Mullen 2011). The findings regarding *outcomes* for patients who report coercive experiences are not encouraging. Early studies found that perceived coercion was associated with avoidance of mental health services, lower patient satisfaction, and negative attitudes toward treatment (Day et al. 2005; Rogers 1993; Swartz et al. 2003). More recent work has found that higher scores of perceived coercion in involuntary patients are associated with lower treatment satisfaction (Katsakou et al. 2010). Voluntary patients who report coercion tend to have less symptomatic improvement than those who are involuntarily committed, when adjusted for higher baseline symptoms and diagnosis, and are more likely to consider their treatment ineffective (Fiorillo et al. 2012; Kallert et al. 2011b).

There is also interesting data concerning patient perspectives on the *sources* of unwelcome influence or control. A consistent finding has been that threats of force, or "negative pressure," are strongly associated with the presence of perceived coercion (Bindman et al. 2005; Iversen et al. 2002; O'Donoghue et al. 2014). "Positive pressures," such as persuasion and inducements, are also reported in up to 60 percent of voluntary patients, though without an association with perceived coercion (O'Donoghue et al. 2014). It appears that this dynamic can depend upon the source of the pressure. For example, both negative *and* positive pressures from mental health professionals are associated with higher perceived coercion (Cascardi and Poythress 1997; Lidz et al. 2000). Notice that this raises the possibility that certain kinds of offers or incentives, from certain types of people, may qualify as coercion, at least on this particular operalization. Another key finding is the association between perceived coercion in involuntary and voluntary patients and a reported lack of clear and fair decision-making

processes, or "*procedural justice*" (Cascardi and Poythress 1997; McKenna et al. 2001; O'Donoghue et al. 2014).

There has been a welcome increase of qualitative research on related issues. Voluntary patients who report coercive experiences largely attribute it to a lack of participation in the admission process, including feeling rushed in their decision to enter hospital (Katsakou et al. 2011). Many voluntary patients report that the possibility of involuntary commitment represents a kind of *implicit threat* throughout the admission process. That is, they believe that they will be legally detained if they do not agree to inpatient treatment (Katsakou et al. 2011). This can include discomfort or disagreement with the use of incentives, with the implication that a wide range of interventions and interactions may be tainted by the lingering possibility of detention (Priebe et al. 2010b). In addition, the qualitative literature points to the importance of *relationships, norms*, and *values*. For example, experiences of coercion—or infringements on freedom of choice—have been linked to negative views of the therapeutic relationship (Theodoridou et al. 2012), and broader social expectations have been reported as sources of pressure (Canvin et al. 2013). When queried about the values that are most important to them during treatment, involuntary patients describe freedom of choice, feeling safe, and respect from staff (Valenti et al. 2014).[8]

What should we make of these findings? An initial impression is that coercive pressures are highly prevalent in psychiatric treatment, at least from the perspective of patients (Newton-Howes and Stanley 2012). Here, though, questions might arise about the veracity of patient reports, especially given the strong association between perceived coercion and symptom severity, not to mention the uncertainty expressed by patients concerning their legal status. An interesting aspect of the work on perceived coercion is the tendency for involuntary patients to describe their hospitalization as voluntary, from 33 to 44 percent (Kjellin et al. 2006), while 10–66 percent of voluntary inpatients either think their legal status is involuntary or are unsure as to whether they are actually free to leave hospital (Bindman et al. 2005; Kjellin et al. 2006; Rogers 1993). One possible interpretation of these data is that perceptions of unwelcome pressure are the product of psychopathology. This is consistent with the association with acute symptomatology, as well as the correlation between perceived coercion and reduced insight or awareness of illness (Bindman et al. 2005). There have been efforts to resolve this matter, including chart reviews and staff interviews in an attempt to compare sources of information. The results have been inconclusive, with some studies confirming patient reports (Lidz et al. 1997) and others finding little agreement between sources (Poulsen and

Engberg 2001). Another reading of the data is available, though, which is that patients find psychiatric treatment to be intrinsically threatening. This would help explain the concerns that many voluntary inpatients have about their actual status, if not the inaccurate views that some involuntary patients have about their own legal status.

Now, to be clear, the subjective experience of pressure or lack of control may not be deemed sufficient to rise to the level of voided consent or mitigated responsibility. Again, on Wertheimer's view of coercion, the key question is whether the intervention in question has this degree of moral force, regardless of how a patient perceives or describes their situation. But of course there is a range of other ethically significant considerations beyond those pertaining to informed consent (nor is consent itself an all-or-nothing affair). The experiences reported by many psychiatric patients may not always qualify as coercion-grade interference, on Wertheimer's terms, but they are morally troubling nonetheless. Given the robust findings on perceived coercion, they should be taken seriously. This is also true when considering their relation to clinical outcomes. Reducing perceptions of coercion is likely to improve outcomes, patient satisfaction, and medication adherence. Similar to the recommendations for reducing more objective forms of coercion, this starts with treating symptoms. Researchers have also stressed the importance of the first encounter with patients for perceptions of procedural justice (Cascardi et al. 2000), as well as providing clear, accessible information to voluntary patients about their ability to leave hospital (O'Donoghue et al. 2014).[9]

Ethical guidelines

Beyond the challenge of interpreting the empirical findings, there is also the question of how these results fit into broader ethical frameworks. There have been few, if any, attempts to provide guidelines for the use of coercion and other forms of influence in mental health. There are of course legal standards and regulatory structures that address coercive measures in many jurisdictions. As mentioned, there have also been proposals that provide classification of the varieties of influence and their attendant justificatory burden. But despite these efforts, researchers have typically avoided detailed analyses of the ethics of influence, especially of the more subtle forms that traditionally fall within the categories of manipulation and persuasion. In this section I consider a notable exception, that of Blumenthal-Barby and colleagues (2013). Their account is a

step in the right direction; it captures a number of important issues and thus should prove influential going forward. That said, I offer critical commentary, picking up several strands from the preceding that prompt a few friendly amendments.

Evaluating influence

Blumenthal-Barby and colleagues (2013) present a typology of influences that cuts across traditional categories. Starting with what usually falls under the heading of persuasion, they identify direct *recommendations* (e.g., advising patients to delay discharge after starting a new medication), appeals to patients' *values and goals* (e.g., emphasizing the positive impact treatment can have on parental responsibilities), and appeals to social *norms* (e.g., mentioning what patients in similar situations tend to do). Next are the manipulative influences of *framing information* (e.g., using non-clinical language rather than more technical terms) and offering *incentives* (e.g., making certain privileges on the ward contingent on other behaviors). Finally, there are the more serious influences of concrete *threats* (e.g., threatening to keep a patient hospitalized unless they follow directives) and different forms of *deception*: concealing (e.g., failing to inform patients of the option to return home in favor of a supported placement), misleading (e.g., presenting information in a way that leads patients to have unrealistic assumptions about their prognosis), and lying (e.g., intentionally misidentifying ward staff due to patient mistrust of mental health professionals).

Each of these methods of influence comes with specific ethical recommendations. Shared by most of them is a set of three obligations. First, the patient's *capacity* for autonomous choice must be evaluated, with greater capacity signaling that influence is less justifiable. Second, one's degree of confidence in the *efficacy* of treatment should be assessed, with a lower level of confidence indicating that influence is less justifiable. And third, the greater the *harm* caused by the influence—and treatment—relative to benefit, then the less justifiable. Several forms of influence also have their own specifications. The permissibility of appealing to a patient's values and goals requires knowledge of that patient, and any social norms used must be accurate and devoid of prejudice and stigma. Incentives used to influence must not include entitlements to which the patient has a right. Threats and deception are to be used only after other efforts have failed, if they will prevent imminent clinical harm, or if they are the only way to protect a patient's autonomy from permanent impairment.

With these qualifications in place, Blumenthal-Barby and colleagues have developed useful guidelines for evaluating influence in mental health. The varieties of influence included are not meant to be comprehensive, or entirely independent from each other, nor is their evaluation expected to be anything close to objective or definitive. It advocates a balancing approach, whereby the risks and benefits associated with any particular case drive ethical deliberation. But it is recognized that this may have limits, such as the possibility of holding that threats and deception are *always* impermissible because of autonomy-based claims (although no additional guidance is provided on this matter). The primacy of capacity is another crucial piece of the guidance, an improvement over ethical and legal approaches that prioritize safety/dangerousness.

Several amendments

Revisiting the findings previously discussed, this framework could be improved in several ways. The first is increased recognition that implicit coercion may be at work in many clinical interactions. As we've seen, voluntary patients are often under the impression that they will be legally detained if they do not agree to admit themselves to hospital. Given the prevalence of involuntary commitment, there is every reason for patients to think that this is a real possibility, and, depending on the nature of the clinical encounter and communication therein, for it be taken as a credible threat. The available data primarily pertains to the admission event, but it's not difficult to imagine how this dynamic could persist over the course of an inpatient stay. In fact, implicit threats may pervade many aspects of treatment. This could change the import of ostensibly benign interactions, including persuasive influences such as recommendations and appeals to values or norms. The potential for such implicit coercion, then, may impact patient autonomy in a way that's underappreciated by Blumenthal-Barby and colleagues' account.[10] It also has implications for the efficacy of treatment, given what we know about perceptions of coercion for patient satisfaction and clinical outcomes.

Beyond the possibility that detention hovers threateningly in the background, one might also think that the professional roles of mental health staff entail certain duties and obligations. Indeed, it has been argued that community mental health workers have a duty to care that prohibits certain kinds of conditionalized proposals (Dunn et al. 2012). Drawing on Wertheimer's idea of a preexisting baseline, any proposal that is made contingent on accepting treatment interventions is arguably a threat—not an offer—if what is being

proposed is already part of the standard treatment package. For example, if staff proposes an activity such as leaving the hospital on an outing but only on condition that patients attend group therapy, this is a wrongful threat if the activity is part of standard treatment. This is in line with Blumenthal-Barby and colleagues' requirement that the use of influence shouldn't include incentives to which the patient has a right. But it should be made explicit that the rights and obligations comprising a patient's moral baseline are partially grounded in the professional duties of mental health staff.

An empirical finding that has implications for any proposed set of ethical guidelines is the association between perceived coercion and procedural justice. The data make it plain that both voluntary and involuntary patients are critical of a lack of clear processes, especially during admission to hospital. The general point is that the availability and quality of information matters. As mentioned, there are several ways of interpreting the findings that patients are unclear about their legal status. In some cases it may be that incapacity or lack of insight is to blame. But in others it is just as likely that a lack of clear and transparent processes is to blame. And it may be that the manner in which incentives and threats are presented is a part of this lack of clarity. Thus more subtle forms of manipulative and persuasive influence may have enhanced importance in the mental health context, including recommendations, appeals to values and norms, and the framing of information.

The ethical analysis of these forms of influence has received relatively little attention in mental health; what work there is tends to focus on the use of "nudges" in the field of public health. Blumenthal-Barby and colleagues (2013) are thus right to highlight the framing of information as potentially significant. Most treatments of nudging center on the question of autonomy; it's fair to say that attempts to motivate ethical concern on these grounds have been mostly unsatisfactory (Cratsley 2015; Quigley 2014). But that shouldn't be taken to suggest that concerns don't apply. For example, there is the well-recognized worry that information shaped in certain ways effectively bypasses rational assessment. Indeed, an opacity condition is the very basis of the effectiveness of techniques inspired by the heuristics and biases tradition (in that they operate outside conscious awareness). If one holds that informed consent requires open deliberation between options, however, the influence of choice architecture can become ethically problematic. This worry may have increased significance in the context of mental health given the cognitive vulnerabilities of many patients.

This is not to deny the potential helpfulness of choice architecture. We might very well think that certain circumstances demand interventions of this type,

justifiable in relation to a patient's degree of autonomy, the effectiveness of the treatment, and risk/benefit calculations. But we do need to be as clear as possible about these justifications; a modicum of scrutiny is required. As we've seen, attention to professional roles and duties may also be significant. Consider the fact that the motivation for some of the original work on heuristics and biases was to *improve* decision-making in medical contexts (Kahneman 2011). Choice architectures that merely make use of biases arguably have a different moral status, as they leave the imperfections of reasoning intact. This point has added resonance in a context where professionals have a duty to treat psychiatric symptoms, many of which directly impact reasoning capacities. Simply put, there may be an ethical tension in harnessing heuristics and biases in the pursuit of alleviating conditions associated with cognitive impairment.

Blumenthal-Barby and colleagues recommend a kind of vigilance on the part of mental health professionals, given the ethical risks of influence. This is salutary advice, especially as there are professional duties that include a set of baseline rights and obligations. This includes providing sufficient information and understanding to meet the requirements of informed consent, but may also involve expectations on the part of patients that aspects of their condition will not be used as a source of influence. Finally, as we've seen in the research on perceived coercion, relationships and social norms have ethical significance. The role of mental health professionals, specifically, appears to inform patients' interpretations of whether a proposal is a wrongful threat or a harmless offer during the process of admission to hospital. And again, these dynamics have significance for clinical outcomes. Therefore, efforts should be made to increase the involvement of nonprofessionals in the process of initiating treatment, such as family members, not to mention the possibility of incorporating the input of layperson and peer support.

Conclusion

The use of coercive measures is widespread in the field of mental health, much of which has significant ethical import. Debate continues over how best to define and distinguish different forms of influence, and there are active empirical research programs pursuing a range of related questions. Efforts to develop ethical guidelines for the use of influence should take some of these findings into account. In particular, the aspects of treatment that appear to be especially salient are the implicit threat of involuntary treatment, the lack of transparent

informational processes, and the quality of therapeutic relationships. Of the findings now available, perhaps the most important is just that the use of pressure and influence is a modifiable variable. Recommendations to reduce the use of coercive measures center around more effective treatment of symptoms, better communication, an emphasis on patient rights, increased training of staff, and the possibility of using advance statements. Progress in these areas should help improve patient experiences while also impacting clinical outcomes. None of this is to deny, however, that there are circumstances where the need for immediate treatment takes precedence over other considerations, when coercion is permissible, unavoidable, or even obligatory.

Acknowledgments

I would like to thank the editors, Robyn Bluhm and Şerife Tekin, for helpful comments on earlier versions of this chapter.

Notes

1 Nozick's (1969) detailed criteria are as follows: (1) A aims to get B to choose action X; (2) A communicates a threat to B; (3) A's threat indicates that if B does not do X, then A will do Y, making B's not X-ing less desirable than B's X-ing; (4) A's threat is credible; (5) B does X; (6) and B does X so as to lessen likelihood that A will do Y. Notice that this definition only applies to specific actions (X), and includes both a publicity requirement (2) and a criterion of successful completion (5). For extended discussion, see Anderson (2011).

2 There is of course a longer story here. The key point is that for Wertheimer the baseline is moral, comprised of preexisting rights and obligations. Consider the example of someone offering to save a drowning person but only on condition of payment. While structured as an offer, the drowning person is arguably threatened with the denial of the right to be saved (and the other person's failure to fulfill their obligation to save). In contrast, Nozick also allowed for a "statistical" reading of the baseline, in terms of the "normally expected course of events." For example, a drug addict is offered their daily fix by their usual dealer but only condition that they carry out a physical assault on a third party. Here a coercion claim might be made on the basis of expectations concerning the usual state of affairs (though moral concerns may also apply).

3 Once again there is a longer story that falls outside the scope of this discussion. For instance, while Feinberg (1986) holds that both threats and offers can constitute

coercion, coercive offers made in circumstances not 'deliberately created' by the coercer do not invalidate consent. Thus coercion always has a limiting effect on voluntariness but not all coercion invalidates consent. It's also worth noting that Feinberg's theory may be particularly well suited for the psychiatric context given its emphasis on psychological pressure.

4 There have been similar proposals in the field of public health, such as the Intervention Ladder from the Nuffield Council on Bioethics (2007).

5 This relates to methodological worries that have exercised the field of philosophy more generally; they are not addressed here.

6 Forced medication was also found to have the strongest association with patient disapproval of treatment (McLaughlin et al. 2016).

7 The outcome data on mandated community treatment describe limited clinical benefit (Molodynski et al. 2016). For discussion of the ethics of coercion focused on community mental health, see Gergel and Szmukler (2016).

8 All of these findings are consistent with research on "coercion contexts" (Sjostrom 2016), as well as emerging evidence on "informal coercion" more generally (reviewed in Hotzy and Jaeger 2016).

9 In many jurisdictions regulations governing discharge are sufficiently restrictive to make voluntary admission something less than fully voluntary (Garakani et al. 2014).

10 More recently, Blumenthal-Barby (2014) has offered an account of manipulation that emphasizes the importance of context and relationships, although not directly applied to psychiatry.

References

Anderson, S. (2010), The Enforcement Approach to Coercion, *Journal of Ethics and Social Philosophy*, 5:1–31.

Anderson, S. (2011), Coercion. In *The Stanford Encyclopedia of Philosophy.* E. Zalta, ed. Winter 2011 edition. URL= http://plato.stanford.edu/archives/win2011/entries/coercion/.

Appelbaum, P. S. (2004), Psychiatric Advance Directives and the Treatment of Committed Patients, *Psychiatric Services*, 55:751–763.

Bindman, J., Reid, Y., Szmukler, G., Tiller, J., Thornicroft, G., and Leese, M. (2005), Perceived Coercion at Admission to Psychiatric Hospital and Engagement with Follow-Up, *Social Psychiatry and Psychiatric Epidemiology*, 40:160–166.

Blumenthal-Barby, J. S. (2012), Between Reason and Coercion: Ethically Permissible Influence in Health Care and Health Policy Contexts, *Kennedy Institute of Ethics Journal*, 22:345–366.

Blumenthal-Barby, J. S. (2014), A Framework for Assessing the Moral Status of "Manipulation." In *Manipulation: Theory and Practice*. C. Coons and M. Weber, eds. Oxford: Oxford University Press.

Blumenthal-Barby, J. S., McCullough, L. B., Krieger, H., and Coverdale, J. (2013), Methods of Influencing the Decisions of Psychiatric Patients: An Ethical Analysis, *Harvard Review of Psychiatry*, 21:275–279.

Burns, T., Rugkasa, J., Molodynski, A., Dawson, J., Yeeles, K., Vazquez-Montes M., and Priebe, S. (2013), Community Treatment Orders for Patients with Psychosis (OCTET): A Randomized Controlled Trial, *Lancet*, 381:1627–1633.

Cairns, R., Maddock, C., Buchanan, A., David, A. S., Hayward, P., Richardson, G., and Hotopf, M. (2005), Prevalence and Predictors of Mental Incapacity in Psychiatric In-Patients, *British Journal of Psychiatry*, 187:379–385.

Canvin, K., Rugkasa, J., Sinclair, J., and Burns, T. (2013), Leverage and Other Informal Pressures in Community Psychiatry in England, *International Journal of Law and Psychiatry*, 36:100–106.

Cascardi, M. and Poythress, N. (1997), Correlates of Perceived Coercion during Psychiatric Hospital Admission, *International Journal of Law and Psychiatry*, 20:445–458.

Cascardi, M., Poythress, N. G., and Hall, A. (2000), Procedural Justice in the Context of Civil Commitment: An Analogue Study, *Behavioral Sciences and the Law*, 18:731–740.

Cratsley, K. (2015), Nudges and Coercion: Conceptual, Empirical, and Normative Considerations, *Monash Bioethics Review*, 33:210–218.

Day, J. C., Bentall, R. P., Roberts, C., Randall, F., Rogers, A., Cattell, D., and Power, C. (2005), Attitudes towards Antipsychotic Medication: The Impact of Clinical Variables and Relationships with Health Professionals, *Archives of General Psychiatry*, 62:717–724.

De Jong, M. H., Kamperman, A. M., Oorschot, M., Priebe, S., Bramer, W., Van De Sande, R., and Mulder, C. L. (2016), Intervention to Reduce Compulsory Psychiatric Admissions: A Systematic Review and Meta-Analysis, *JAMA Psychiatry*, 73:657–664.

de Stefano, A. and Ducci, G. (2008), Involuntary Admission and Compulsory Treatment in Europe: An Overview, *International Journal of Mental Health*, 37:10–21.

Dunn, M., Maughan, D., Hope, T., Canvin, K., Rugkasa, J., Sinclair, J., and Burns, T. (2012), Threats and Offers in Community Mental Healthcare, *Journal of Medical Ethics*, 38:204–209.

Faden, R. R. and Beauchamp, T. L. (1986), *A History and Theory of Informed Consent*, Oxford: Oxford University Press.

Feinberg, J. (1986), *Harm to Self*, New York: Oxford University Press.

Fiorillo, A., De Rosa, C., Del Vecchio, V., Jurjanz, L., Schnall, K., Onchev, G., and Kallert, T. (2011), How to Improve Clinical Practice on Involuntary Hospital Admissions of Psychiatric Patients: Suggestions from the EUNOMIA Study, *European Psychiatry*, 26:201–207.

Fiorillo, A., Giacco, D., De Rosa, C., Kallert, T., Katsakou, C., Onchev, G., and Priebe, S. (2012), Patient Characteristics and Symptoms Associated with Perceived Coercion during Hospital Treatment, *Acta Psychiatrica Scandivica*, 125:460–467.

Garakani, A., Shalenberg, E., Burstin, S. C., Brendel, R. W., and Appel, J. M. (2014), Voluntary Psychiatric Hospitalization and Patient-Driven Requests for Discharge: A Statutory Review and Analysis of Implications for the Capacity to Consent to Voluntary Hospitalization, *Harvard Review of Psychiatry*, 22:241–249.

Gardner, W., Hoge, S. K., Bennett, N. S., Roth, L., Lidz, C., Monahan, J., and Mulvey, E. (1993), Two Scales for Measuring Patients' Perceptions of Coercion during Mental Hospital Admission, *Behavioral Sciences and the Law*, 11:307–321.

Gergel, T. and Szmukler, G. (2016), The Ethics of Coercion in Community Mental Health Care. In *Coercion in Community Mental Health Care: International Perspectives*. A. Molodysnki, J. Rugkasa, and T. Burns, eds. Oxford: Oxford University Press.

Hawkins, J. S. and Emanuel, E. J. (2005), Clarifying Confusions about Coercion, *The Hastings Center Report*, 35:16–19.

Hoge, S. K., Lidz, C. W., Eisenberg, M., Gardner, W., Monahan, J., Mulvey, E., and Bennett, N. (1997), Perceptions of Coercion in the Admission of Voluntary and Involuntary Psychiatric Patients, *International Journal of Law and Psychiatry*, 20:167–181.

Hotzy, F. and Jaeger, M. (2016), Clinical Relevance of Informal Coercion in Psychiatric Treatment—A Systematic Review, *Frontiers in Psychiatry*, 7: 197.

Hoyer, G., Kjellin, L., Engberg, M., Kaltiala-Heino, R., Nilstun, T., Sigurjonsdottir, M., and Syse, A. (2002), Paternalism and Autonomy: A Presentation of a Nordic Study on the Use of Coercion in the Mental Health Care System, *International Journal of Law and Psychiatry*, 25:93–108.

Husum, T. L., Bjorngaard, J. H., Finset, A., and Ruud, T. (2010), A Cross-Sectional Prospective Study of Seclusion, Restraint and Involuntary Medication in Acute Psychiatric Wards: Patient, Staff and Ward Characteristics, *BMC Health Services Research*, 10:89.

Iversen, K. I., Høyer, G., Sexton, H., and Gronli, O. K. (2002), Perceived Coercion among Patients Admitted to Acute Wards in Norway, *Nordic Journal of Psychiatry*, 56:433–439.

Jaeger, M., Ketteler, D., Rabenschlag, F., and Theodoridou, A. (2014), Informal Coercion in Acute Inpatient Setting—Knowledge and Attitudes Held by Mental Health Professionals, *Psychiatry Research*, 220:1007–1011.

Kahneman, D. (2011), *Thinking Fast, Thinking Slow*, New York: Penguin.

Kalisova, L., Raboch, J., Nawka, A., Sampogna, G., Cihal, L., Kallert, T. W., and Fiorillo, A. (2014), Do Patient and Ward-Related Characteristics Influence the Use of Coercive Measures? Results from the EUNOMIA International Study, *Social Psychiatry and Psychiatric Epidemiology*, 49:1619–1629.

Kallert, T. W., Glockner, M., Onchev, G., Raboch, J., Karastergiou, A., Solomon, Z., and Kjellin, L. (2005), The EUNOMIA Project on Coercion in Psychiatry: Study Design and Preliminary Data, *World Psychiatry*, 4:168–172.

Kallert, T. W., Mezzich, J. E., and Monahan, J. (2011a), *Coercive Treatment in Psychiatry: Clinical, Legal and Ethical Aspects*, West Sussex: Wiley.

Kallert, T. W., Katsakou, C., Adamowski, T., Dembinskas, A., Fiorillo, A., Kjellin, L., and Priebe, S. (2011b), Coerced Hospital Admission and Symptom Change—A Prospective Observational Multi-Centre Study, *PLoS One*, 6:e28191.

Katsakou, C. and Priebe, S. (2006), Outcomes of Involuntary Hospital Admission—A Review, *Acta Psychiatrica Scandinavica*, 114:232–241.

Katsakou, C., Bower, L., Amos, T., Morriss, R., Rose, D., Wykes, T., and Priebe, S. (2010), Coercion and Treatment Satisfaction among Involuntary Patients, *Psychiatric Services*, 61:286–292.

Katsakou, C., Marougka, S., Garabette, J., Rost, F., Yeeles, K., and Priebe, S. (2011), Why Do Some Voluntary Patients Feel Coerced into Hospitalization? A Mixed Methods Study, *Psychiatry Research*, 187:275–282.

Keown, P., Weich, S., Bhui, K. S. and Scott, J. (2011), Association Between Provision of Mental Illness Beds and Rage of Involuntary Admissions in the NHS in England 1988–2008: Ecological Study, *British Medical Journal*, 343:d3736.

Kjellin, L., and Wallsten, T. (2010), Accumulated Coercion and Short-Term Outcome of Inpatient Psychiatric Care, *BMC Psychiatry*, 10:1–7.

Kjellin, L., Hoyer, G., Engberg, M., Kaltiala-Heino, R., and Sigurjonsdottir, M. (2006), Differences in Perceived Coercion at Admission to Psychiatric Hospitals in the Nordic Countries, *Social Psychiatry and Psychiatric Epidemiology*, 41:241–247.

Largent, E., Grady, C., Miller, F. G., and Wertheimer, A. (2013), Misconceptions about Coercion and Undue Influence: Reflections on the Views of IRB Members, *Bioethics*, 27:500–507.

Lidz, C. W., Mulvey, E. P., Hoge, S. K., Kirsch, B. L., Monahan, J., Bennett, N. S., and Roth, L. H. (1997), The Validity of Mental Patients' Accounts of Coercion in the Hospital Admission Process, *Law and Human Behavior*, 21:361–376.

Lidz, C. W., Mulvey, E. P., Hoge, S. K., Kirsch, B., Monahan, J., Bennett, N., and Roth, L. (2000), Sources of Coercive Behaviours in Psychiatric Admissions, *Acta Psychiatrica Scandinavica*, 101:73–79.

McKenna, B. G., Simpson, A. I. F., Coverdale, J. H., and Laidlaw, T. M. (2001), Analysis of Procedural Justice during Psychiatric Hospital Admission, *International Journal of Law and Psychiatry*, 24:573–581.

McLaughlin, P., Giacco, D., and Priebe, S. (2016), Use of Coercive Measures during Involuntary Psychiatric Admission and Treatment Outcomes: Data from a Perspective Study across 10 European Countries, *PLoSONE*, 11:e0168720.

McMillan, J. (2010), Coercive Offers and Research Participation: A Comment on Wertheimer and Miller, *Journal of Medical Ethics*, 36:383–384.

Molodynski, A., Turnpenny, L., Rugkasa, J., Burns, T., and Moussaoui, D. (2014), Coercion and Compulsion in Mental Healthcare—An International Perspective, *Asian Journal of Psychiatry*, 8:2–6.

Molodynski, A., Rugkasa, J. and Burns, T., eds. (2016), *Coercion in Community Mental Health Care: International Perspectives*, Oxford: Oxford University Press.

Newton-Howes, G. and Mullen, R. (2011), Coercion in Psychiatric Care: Systematic Review of Correlates and Themes, *Psychiatric Services*, 62:465–470.

Newton-Howes, G. and Stanley, J. (2012), Prevalence of Perceived Coercion among Psychiatric Patients: Literature Review and Meta-Regression Modelling, *The Psychiatrist*, 36:335–340.

Nozick, R. (1969), Coercion, pp. 440–472. In *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*. S. Morgenbesser, P. Suppes, and M. White, eds. New York: St. Martin's Press.

Nuffield Council on Bioethics (2007), *Public Health: Ethical Issues*, London: Nuffield Council on Bioethics.

O'Donoghue, B., Roche, E., Shannon, S., Lyne, J., Madigan, K., and Feeney, L. (2014), Perceived Coercion in Voluntary Hospital Admission, *Psychiatry Research*, 215:120–126.

Poulsen, H. D. and Engberg, M. (2001), Validation of Patients' Statements on Coercive Measures, *Acta Psychiatrica Scandinavica*, 103: 60–65.

Priebe, S., Katsakou, C., Amos, T., Leese, M., Morriss, R., Rose, D., and Yeeles, K. (2009), Patients' Views and Readmissions 1 Year after Involuntary Hospitalization, *British Journal of Psychiatry*, 194:49–54.

Priebe, S., Katsakou, C., Glockner, M., Dembinskas, A., Fiorillo, A., Karastergiou, A., and Kallert, T. (2010a), Patients' Views of Involuntary Hospital Admission after 1 and 3 Months: Prospective Study in 11 European Countries, *British Journal of Psychiatry*, 196:179–185.

Priebe, S., Sinclair, J., Burton, A., Marougka, S., Larsen, J., Firn, M., and Ashcroft, R. (2010b), Acceptability of Offering Financial Incentives to Achieve Medication Adherence in Patients with Severe Mental Illness: A Focus Group Study, *Journal of Medical Ethics*, 36:463–468.

Quigley, M. (2014), Are Health Nudges Coercive? *Monash Bioethics Review*, 32:141–158.

Raboch, J., Kalisova, L., Nawka, A., Kitzlerova, E., Onchev, G., Karastergiou, A., Magliano, L., Dembinskas, A., Kiejna, A., Torres-Gonzales, F., Kjellin, L., Priebe, S., and Kallert, T. W. (2010), Use of Coercive Measures during Involuntary Hospitalization: Findings from Ten European Countries, *Psychiatric Services*, 61:1012–1017.

Radden, J. (2015), Unique Ethical Challenges for Psychiatric Practice. In *Oxford Handbook of Psychiatric Ethics*. J. Sadler, W. Van Staden, and K. W. M. Fulford, eds. Oxford: Oxford University Press.

Rogers, A. (1993), Coercion and "Voluntary" Admission: An Examination of Psychiatric Patient Views, *Behavioral Sciences and the Law*, 11:259–267.

Singh, S. P., Burns, T., Tyrer, P., Islam, Z., Parsons, H., and Crawford, M. J. (2014), Ethnicity as a Predictor of Detention under the Mental Health Act, *Psychological Medicine*, 44:997–1004.

Sjostrom, S. (2016), Coercion Contexts—How Compliance Is Achieved in Interaction. In *Coercion in Community Mental Health Care: International Perspectives*. A. Molodynski, J. Rugkasa, and T. Burns, eds. Oxford: Oxford University Press.

Swartz, M. S., Swanson, J. W., and Hannon, M. J. (2003), Does Fear of Coercion Keep People Away from Mental Health Treatment? Evidence from a Survey of Persons with Schizophrenia and Mental Health Professionals, *Behavioral Sciences & Law*, 21:459–472.

Szmukler, G. (2015), Compulsion and "Coercion" in Mental Health Care, *World Psychiatry*, 14:259–261.

Szmukler, G. and Appelbaum, P. S. (2008), Treatment Pressures, Leverage, Coercion, and Compulsion in Mental Health Care, *Journal of Mental Health*, 17:233–244.

Theodoridou, A., Schlatter, F., Ajdacic, V., Rossler, W., and Jager, M. (2012), Therapeutic Relationship in the Context of Perceived Coercion in a Psychiatric Population, *Psychiatry Research*, 200:939–944.

Valenti, E., Giacco, D., Katsakou, C., and Priebe, S. (2014), Which Values Are Important for Patients during Involuntary Treatment? A Qualitative Study with Psychiatric Patients, *Journal of Medical Ethics*, 40:832–836.

Wertheimer, A. (1987), *Coercion*, Princeton, NJ: Princeton University Press.

Wertheimer, A. (1993), A Philosophical Examination of Coercion for Mental Health Issues, *Behavioral Sciences and the Law*, 11:239–258.

Wertheimer, A. and Miller, F. (2008), Payment for Research Participation: A Coercive Offer? *Journal of Medical Ethics*, 34:389–392.

Wykes, T., Csipke, E., Williams, P., Koeser, L., Nash, S., Rose, D., and McCrone, P. (2017), Improving Patient Experiences of Mental Health Inpatient Care: A Randomized Controlled Trial, *Psychological Medicine*, 48:488–497.

15

Voice, Silencing, and Listening Well: Socially Located Patients, Oppressive Structures, and an Invitation to Shift the Epistemic Terrain

Nancy Nyquist Potter

Introduction

One night, a thirty-year-old person of color, who I'll call "Eleanor," visited emergency psychiatry, where she previously had been seen a number of times before. The file identified the patient as male; however, she requested that she be referred to with the female pronoun, as she is transgender. She reported feelings of depression and anxiety, and presented to emergency psychiatry with suicidal ideation. This patient repeatedly expressed that she feared for her safety in the shelters, while the attending and the resident continued to refer to the patient by male pronouns even though she had asked to be addressed as female. The health care team believed that the patient thought hospitalization was a way to avoid spending another night in a shelter; they assessed her as not being a danger to herself and not needing crisis intervention. Therefore, she was discharged.

This chapter addresses the ways that people in mental distress experience themselves as silenced or otherwise not heard. It is not an anti-psychiatry essay; I believe that mental illnesses exist and that clinicians often do significant good in working with patients. Nevertheless, I will argue that there is a pressing need for the epistemic and ethical terrain to shift between clinicians and their patients/service users. This shift is relational and implicates both or all parties to the clinical encounter, although those with more relational and professional power bear more of the responsibility to make such alterations possible.[1] To make sense of this argument, I believe that an understanding of democratic values and of mechanisms of oppression is useful. The next sections, therefore,

provide a more general geography—a geography of politics and economics— for the particular landscape in which clinical navigation is required. The section Defining Structural Violence and Oppression explains how societal trends toward neoliberalism affect health care, and mental health care, in particular; the section Mental Illness and Mechanisms of Oppression analyzes several concepts from political and moral philosophy to explain how the distribution of goods and resources affects members of different group in unequal ways, and next two sections apply these concepts to psychiatry. The concluding section ends by making some general comments about psychiatry given the preceding sections.

Enough with the metaphors! Democracy for real

Presumably, an ideal political landscape would enable us to live well. Political theorists thus spill much ink (well, cyberspace) to develop values, policies, and structures that would provide the groundwork for an ideally just and democratic society in which we flourish. That the two prevailing values of democracy are said to be liberty and equality will come as no surprise to twentieth- and twenty-first-century scholars of Rawlsian conceptions of justice. However, liberty seems to demand that we give people maximum decision-making freedom regarding their own money, while equality calls for bridging economic gaps between different social groups. It is widely recognized that these two values are in tension to the degree that, although there is an abundance of literature on politics, policy, and economics to sort this out, the most satisfactory response, to my mind, is to understand democratic values as contestable (Mouffe 1996). In particular, I think it is dangerous to uphold any political theory and praxis that encloses and maintains a regime. In particular, liberal democracy gets some of its endurance from a misguided and damaging individualism. By this I mean that the Western world's abiding commitment to seeing ourselves as autonomous individuals who will thrive in an environment of non-interference (serving to the ideal of liberty) misses the moral mark of responsibilities to vulnerable others, to moral requirements to care for the indigent and immigrant populations, and to a pluralist and inclusive communicative process (serving to the ideal of equality). Yet the words "inclusive" and "pluralist" have no teeth if they cannot be realized within sociopolitical and interpersonal relations. What constitutes democracy (so to speak) depends on context and intention. As Wendy Brown says:

> "Democracy" is among the most contested and promiscuous terms in our modern political vocabulary. In the popular imaginary, "democracy" stands for everything from free elections to free markets, from protests against dictators to law and order, from the centrality of rights to the stability of states, from the voice of the assembled multitude to the protection of individuality and the wrong of dicta imposed by crowds … Democracy comes in so many varieties—social, liberal, radical, republican, representative, authoritarian, direct, participatory, deliberative, plebiscite—that such claims often speak past one another. (Brown 2015: 19)

Instead of discussing historical shifts in understanding democracy, I want to mark the direction it has turned: to what is, in some circles, called **neoliberalism.** Neoliberalism is an ideology with serious pragmatic consequences: it holds that private markets are more cost-effective and consumer friendly than a national and public market. It is of grave concern to health care in that it treats health care as a private good for sale rather than a public good that is paid for with tax dollars (McGregor 2011). The reason this is so crucial to understand is that it has significant effects not only on access and care for the mentally ill, but also on how we conceptualize people with mental illnesses. (I'll say more about this shortly.) This form of reasoning is ubiquitous today in government, the workplace, jurisprudence, education, culture, and a vast range of everyday life that involves "converting the distinctly political character, meaning, and operation of democracy's constituent elements into economic ones" (Brown 2015: 17). Since 9/11, capitalism increasingly has been linked to the notion of liberty, but that liberty consists in freedom from all constraints, an economic "system that has let itself go—that no longer has to work to keep us as customers, that can be as antisocial, antidemocratic and boorish as it wants" (Klein 2007: 319). Naomi Klein thus describes neoliberalism as a "commitment to the policy trinity—the elimination of the public sphere, total liberation for corporations, and skeletal social spending" (Klein 2007: 17). Rodolpho Walsh likens the crimes against humanity committed during the Argentinean dictatorship[2] to the local manifestation of neoliberalism as "planned misery," whereby

> it is in the economic policy of this government where one discovers not only the explanation of the crimes, but a greater atrocity which punishes millions of human beings through planned misery ….You only have to walk around greater Buenos Aries for a few hours to check the speed with which such a policy transforms the city into a 'shantytown' of ten million people. (cited in Klein 2007: 117)

"Planned misery," in fact, may describe current attempts in the United States to repeal Obamacare, leaving millions of Americans without health insurance. Klein suggests that a better name for this system is not liberal, conservative, or capitalist, but corporatist.

Brown, like Mouffe, advocates an open and contestable understanding of democracy, yet Brown nevertheless insists on its value in connoting self-rule by the people (leaving open who the "we" are but expecting that it embraces inclusivity and plurality). The idea is that it is important to preserve an opposition to tyranny, fascism, totalitarianism, and corporatism as well as a conceptual and political stance against a transmutation of the rule of the people into governmentality of neoliberal reasoning (Brown 2015: 20). Referring to President Obama's State of the Union address in 2013, Brown says even of his speech that "the state's table of purposes and priorities has become indistinguishable from that of modern firms, especially as the latter increasingly adopts concerns with justice and sustainability" (Brown 2015: 27). This fusing of the state with the firm or corporation raises questions about the fate of "rule by and for the people when neoliberal reason configures both soul and city as contemporary firms, rather than polities" (Brown 2015: 27).

Following Brown and Klein, I point to ethical concerns entailed in the rise of neoliberalism's affirmation of free (but private) markets as the *summum bonum* of the good life. The two politico-ethical issues of particular concern are the likelihood of intensified inequality, and the commercialization of goods that ought not to be turned into commodities, such as health care. The effects of neoliberalism will turn out to be directly relevant to clinician–patient relations.

Defining structural violence and oppression

As the previous section indicates, the ideals of equality, liberty, inclusivity, and pluralism that are upheld only in *theoretical* democracy are under threat. They also are out of reach and out of touch with real people's lives with struggle, discrimination, hatred, and danger. Lisa Tessman argues that the actual world is one where many (if not most) people live in adverse conditions, including "the more wretched conditions present under some forms of oppression" (Tessman 2009: 48). Ideal theories—ones that uphold a standard for the perfectly just or moral society—often function as ideologies that, while claiming to be applicable to all, are out of reach and unattainable for many, if not most, people. Shane O'Neill likewise argues that "the social hopes that should be given priority are

those of people who are suffering the brutalizing effects of the worst structural injustices" (O'Neill, 2010: 130). He suggests that, instead of trying to develop and put into practice a fully just state, we ought to "give effective voice to people who feel humiliated and vulnerable because of the oppression they endure or the extreme poverty that is imposed on them" (O'Neill, 2010: 130).

Following Tessman's urging that thinking about how to be moral requires that we consider real people in our actual lives and not an idealized and, typically, generic version of what it means to be a moral agent, I will explain the mechanism of oppression. To begin, I employ the concept of structural violence. **Structural violence** refers to social arrangements that place members of certain groups in harm's way. "The arrangements are *structural* because they are embedded in the political and economic organization of our social world; they are *violent* because they cause injury to people (typically, not those responsible for perpetuating such inequalities)" (Farmer et al. 2006). One kind of structural violence is found in the distribution of access and resources in health care, where the poor, minorities, refugees, undocumented workers, and other marginalized groups frequently have little to no health care because they cannot afford it, and in a neoliberal society, it is not provided as a public good. It is especially important to grasp the effects of structural violence as it correlates with neoliberalism in three ways: First, people living with mental illnesses cannot afford the care and medications they need; second, psychiatrists are encouraged to medicate patients with the newest (and often most expensive) drugs; and universities constrain the amount of time that psychiatrists can spend with their patients. Thus, a group of people already stigmatized and marginalized—the mentally ill—receives poorer care. Looking at this problem by means of the concept of oppression will further clarify the issues at stake.

Oppression is a particular kind of harm that is interwoven with structural violence. It is done to members of groups because they are, or are perceived to be, members of those groups. The harms of oppression are all-encompassing: they not only affect agency and freedom of movement but also people's safety, their health, their sense of self, their relational experiences, their material opportunities, bodily comportment, and their relationships to other groups. The harms of oppression may be passed down through generations, through what is called transgenerational trauma (Potter 2015). For example, the trauma that Africans experienced of being kidnapped and sold in auctions to white masters who often beat them, treated them like animals, and raped the women, is passed on through the anxiety, grief, suppressed anger, and fear of punishment or death that is and kept alive through ongoing racism. These are harms to take seriously.

The definition and conceptualization of oppression I will use comes from Ann Cudd. She states, "Oppression means a harm through which groups of persons are systematically and unfairly or unjustly constrained, burdened, or reduced by any of several forces. Oppression is a normative concept that names a *social injustice*. Oppression is always wrong" (Cudd 2006: 23; emphasis in original). By identifying oppression as a social injustice, Cudd frames it as a mechanism that operates on social groups and by social groups through institutions, practices, laws, and norms. That is, oppression is socially structured; it forms a complex matrix of scaffoldings by which social groups are confined, limited, disciplined, and shaped. "For every social group that is oppressed, there are correlative social groups whose members benefit, materially or psychologically, from the oppression" (Cudd 2006: 25).

But even this conceptual framework elides the complex experiences of many people's lives. Black feminists have long pointed out that a full understanding of oppression requires the recognition that race, class, and gender intersect and, hence, that oppressions are overlapping; no simple formula for identifying oppressors and oppressed exists. When we ignore this, the experiences of black women and other women of color are distorted. As a way for us to talk about people's phenomenological experiences of marginalization and erasure, Kimberlé Crenshaw introduced the metaphor of intersectionality (Crenshaw 1991). **Intersectionality** is not synonymous with oppression; it is a metaphor, and sometimes a research methodology, for understanding and uncovering multi-valenced oppressions.

> Consider an analogy to traffic in an intersection, coming and going in all four directions. Discrimination, like traffic through an intersection, may flow in one direction, and it may flow in another. If an accident happens in an intersection, it can be caused by cars traveling from any number of directions and, sometimes, from all of them. Similarly, if a Black woman is harmed because she is in the intersection, her injury could result from sex discrimination or race discrimination. (Crenshaw 1989: 149)

Crenshaw remarks that it is not easy to disentangle the skid marks to determine an explanation for the accident; similarly, it is not easy—or even possible or desirable—to understand overlapping oppressions by trying to separate the "gender" part from the "race" part: living intersectional lives is structural, political, and representational (Carastathis 2014).

> Oppression and privilege by race, ethnicity, gender, sexual orientation, class, nationality, and so on, do not act independently of one another in our individual

> lives or in our social structures; instead, each kind of oppression or privilege is shaped by and works through the others. These compounded, intermeshed systems of oppression and privilege in our social structures help to produce (a) our social relations; (b) our experiences of our own identity; and (c) the limitations of shared interests even among members of "the same" oppressed or privileged group. (Garry 2012)

While the origins of the concept are rooted in Black feminist thought, intersectionality also reminds us that other oppressive conditions tend to be neglected, such as that of genderqueer people, as Eleanor in the above example, or those living with disabilities. It may not be obvious how an understanding of oppression is relevant to the lives of people living with mental illnesses, so let me say a bit more. As James Charlton says, "the lived oppression that people with disabilities have experienced and continue to experience is a human rights tragedy of epic proportions" (Charlton 2000: ix). This is a powerful example of oppressions at work. Amartya Sen reports that 1 in 10 people live with physical and mental disabilities and that these are among the most deprived human beings as well as the most neglected (Sen 2009). This oppression is rooted in material, political, social, economic and cultural dimensions of everyday life and results in poverty, isolation, and indignity (Charlton 2000: ix).

A broader point is that the ontological status of human animals is not that of natural "pure" entities: we are all born into and, hence, subjected to distributions of power and disadvantage that form the scaffolding upon which we perceive and are perceived and out of which we labor in the material of the world to shape it more or less to our expectations. We are not "just" an individual human, or a generic "self," or even a particular sex. We move between and among people whose lived experiences are fraught with structural oppressions; and we are those people. It seems clear, too, that some of those with whom we move, or move away from, constitute us in ways we do not accept or that are hostile to us. Maria Lugones frames the ontological and phenomenological experience of "self" in terms of "worlds" and "**world**"**-traveling**.

> For something to be a "world" in my sense it has to be inhabited at present by some flesh and blood people. That is why it cannot be a utopia. It may also be inhabited by some imaginary people. It may be inhabited by people who are dead or people that the inhabitants of this "world" met in some other "world" and now have in this "world" in imagination. A "world" in my sense may be an actual society given its dominant culture's description and construction of life, including a construction of the relationships of production, of gender, race, etc. … But a "world" can also be such a society given a non-dominant construction,

> or it can be such a society or *a* society given an idiosyncratic construction. As we will see, it is problematic to say that these are all constructions of the same society. But they are different "worlds." (Lugones 1987: 9–10; emphasis in original)

Lugones writes that she is aware that, in some worlds, she seems to have an attribute (for her, playfulness) that she does not have in other worlds, worlds that may be hostile to her. And in some worlds, she is taken up in a way that she does not recognize. She can be constructed as *stereotypically* latin, or as latin; they are different simultaneous constructions of persons that are part of different "worlds" (Lugones 1987: 11). The experience of shifting from one "world" to another she calls "traveling." I will return to this point at the end of the chapter, but for now, it is enough to understand that people who do this traveling experiences themselves as being different in different worlds.

The presence of "worlds" and intersections of oppression are manifested in the distribution of materiality, politics, law, sociality, education, and safety. But they are also experienced intra-psychically, in terms of internalized oppression: when we take the representations and images of our oppressor and use them against ourselves. As Cudd explains, "Psychological oppression occurs when one is oppressed through one's mental states, emotionally or by manipulation of one's belief states, so that one is psychologically stressed, reduced in one's own self-image, or otherwise psychically harmed" (Cudd 2006). Oppression thus names a circumstance in which four conditions are satisfied:

1. *The harm condition*: There is a harm that comes out of an institutional practice.
2. *The social group condition*: The harm is perpetrated through a social institution or practice on a social group whose identity exists apart from the oppressive harm in (1).
3. *The privilege condition*: There is another social group that benefits from the institutional practice in (1).
4. *The coercion condition*: There is unjustified coercion or force that brings about the harm. (Cudd 2006: 25)

To summarize so far: democratic values are in tension both with the rise of neoliberalism, that increasingly is marketing freedom in the guise of corporatism, and with the heightened inequalities and injustices that entrench, rather than liberate, oppressed people's agency and autonomy. In the next section, I turn to two of the mechanisms of oppression: stereotyping and microaggressions. These concepts lays additional groundwork for understanding how the lives of people

living with mental illnesses are affected by social conditions and, in particular, how such mechanisms can seep into clinician–patient relations.

Mental illness and mechanisms of oppression

Typically, clinicians are not trained to understand the social structures that shape people's lives—including their own—and, furthermore, are not trained in ways that would make it possible for them to address systemic problems. There are many reasons for this, including the institutional structures of medicine and the epistemic commitments in psychiatry and psychology (cf Potter 2016), but I will focus on two of them and then explain how they are relevant to clinical work. First, stereotypes: stereotypes play a crucial role in how individuals think about, feel about, and decide to act toward others (Cudd 2006: 68). Ann Cudd defines them as "generalizations that we make about persons based on characteristics that we believe they share with some identifiable group" (Cudd 2006: 69). Thus, the formation of stereotypes is a type of categorizing. The cognitive act of stereotyping involves a complex series of inferences about characteristics that individuals believe set people in one group apart from people in other groups; as such, they form the foundation of our beliefs about groups (Cudd 2006). Most of us understand stereotypes primarily by their negative connotations. A negative stereotype is a false or misleading generalization that is intransigent with respect to evidence to the contrary. It is a generalizing, fixed, false belief. Stereotypers are "cognitive misers"; they are efficient, intractable, and often wrong, generalizers. In other words, stereotypes do not only help us make quick and efficient judgments; they buffer us from anxieties and fears by creating an external Other who is "bad" to our "good," and they arise when we perceive that our sense of self-integration is threatened (Gilman 1985: 15–35).

Studies in cognitive psychology give us reason to think that stereotyping is an unavoidable feature of cognition. But, as Cudd points out, much injustice is done to individuals (and to groups) with the application and reproduction of stereotypes in our attitudes and treatment of one another. We tend to favor in-groups and disadvantage out-groups of the stereotype while simultaneously creating and maintaining those very in- and out-groups. The injustice done through stereotyping suggests that stereotyping activity is immoral (Cudd 2006: 69). For example, Patricia Hill Collins identifies the persistent images of Black women as mammies, matriarchs, welfare mothers, and hot mamas. Collins calls such representations "controlling images" and argues that they

serve to naturalize racism, sexism, and poverty and make them appear to be an inevitable part of everyday life (Collins 2000: 69). Or consider the stereotyping of Muslim Americans, a pattern of representation that affects both interactions and immigration. Muslim Americans are characterized as terrorists who may be required to register with government databases. "While appeals to the media for accuracy and fairness continue, newspaper headlines regularly print the words 'Islam' and 'Muslim' next to words like 'fanatic,' 'fundamentalist,' 'militant,' 'terrorist' and 'violence.' Uses of the term 'jihad' in television programs and films are also illustrative" (ThePluralismProject.com). As the Trump presidency took office in 2017, a Registry akin to that of registries for Japanese Americans or for Jews in 1930s nationalist Germany was planned to keep track of "legitimate" immigrants and to be able to separate the acceptable from the "dangerous." As Peter Burns and James Gimpel explain, "Those who believe immigrants to be unintelligent, dirty, unwilling to learn English, [religious fanatics, dangerous,] or unwilling to work as hard as natives [*sic*] are more likely to oppose immigration" (Burns and Gimpel 2000: 204).

Stereotypes about the mentally ill abound, and play a central role in how people conceptualize and respond to those living with (or just perceived to be living with) mental illness. As Patrick Corrigan comments, "Commonly held stereotypes about people with mental illness include violence (people with mental illness are dangerous), incompetence (they are incapable of independent living or real work), and blame (because of weak character, they are responsible for the onset and continuation of their disorder)" (Corrigan 2004: 319). People with schizophrenia are deranged and dangerous, those with depression need to pull themselves up by the bootstraps, people with borderline personality disorder are manipulative and attention-seeking, and so on. Such stereotypes seem to legitimize decreasing or eliminating pre-conditions for people with mental illnesses on the grounds that they place an undue burden to the health care system.

Another mechanism of oppression is that of microaggressions. Microaggressions are "brief and commonplace daily verbal, behavioral, or environmental indignities, whether intentional or unintentional, that communicate hostile, derogatory, or negative slights and insults toward members of marginalized groups" (adapted from Sue et al. 2007). One kind of microaggression is that of microinsults, where statements or actions indirectly belittle a marginalized person. For example, a microinsult might be when a clinician calls in the parent of an adult, competent patient and then mostly talks with the parent; the patient may feel infantilized or treated as an

unreliable or trustworthy source. Another kind of microaggression is that of microinvalidations; this occurs with statements and behaviors that negate or nullify a marginalized person's experiences or realities. An example of this is when the clinician suggests to the patient that "underneath her anger (or her defiance) is …." thereby treating that anger or that defiance only as a symptom, or a primitive defense mechanism, rather than something to investigate that might have reasonableness and thus a need to be given uptake.

Returning to the opening case of Eleanor, I suggest that stereotypes and unconscious microaggressions that the health care team tapped into in emergency psychiatry operated to entrench Eleanor's lived experiences of gendered, racialized, and class-based (i.e., her homelessness) oppression at every turn. She was interpreted, it would seem, as not really trans, exploiting race and sexual identity politics in an attempt to manipulate the health care system, and malingering. The result was that she experienced a profound sense of loneliness—a feeling that the health care team contributed to. In the next section, I will argue that epistemic problems of silence and voice collide with democratic values and reproduce oppressive structures.

Silence and voice

I have suggested that structural violence and oppressive forces reproduce distorted representations, values, interpretive schemas, epistemic authority, bodies, and meaning-making. One effect of the power of these intertwined aspects of many people's everyday lives is that communication itself is embedded in patterns of structural injustice (cf. Kukla 2014; Potter 2000). Communication is an interdependent phenomenon, by which I mean that both speakers and listeners are dependent on one another for communication to be effective. Speakers depend on audiences' understanding of the illocutionary act for their communication to be successful (cf. Kukla 2014; Langton 1993), while audiences depend on speakers to give them good information and to be trustworthy (cf. Grasswick 2012). Communication thus is intertwined with epistemic power and distribution of knowers and not-knowers. By not-knowers, I do not mean people who do not know about a particular topic, but people who are not counted as being even prima facie knowers. Recent work in epistemology has analyzed ways that issues of credibility, voice, and silencing/silence follow along an arch of hegemonic relations. Specifically, reliance on others' testimony is so basic to everyday life that

we often overlook the ways it operates. As Miranda Fricker (2007) suggests, assessments of credibility follow social patterns of structural domination and subordination, whereby people in minority groups are considered less credible both because they are viewed as incapable of being knowers and because their character is untrustworthy. So, their testimony is considered incredible and discounted.

This is a mechanism of silencing. Rae Langton explains it this way:

> If you are powerful, you sometimes have the ability to silence the speech of the powerless. One way might be to stop the powerless from speaking at all. Gag them, threaten them, condemn them to solitary confinement. But there is another, less dramatic but equally effective, way. Let them speak. Let them say whatever they like to whomever they like, but stop that speech from counting as an action … Some kinds of speech acts are unspeakable for women in some contexts; although the appropriate words can be uttered, those utterances fail to count as the actions they were intended to be. (Langton 1993: 299)

Women's communications are discounted differently depending on one's social locations and context and, as I have argued, we must understand the failure to be given uptake as occurring intersectionally, and not just in terms of gender, or race, or class, as if those are independent category experiences. Being silenced, and having the power to silence others, is oppressive in several ways: materially, psychologically, politically, and epistemically. Recall the opening case of Eleanor; the health care team's dismissal of her concerns meant that she was left unsafe either in a shelter or on the street, the team caused her anguish in that she felt invisible to the others and experienced herself as unheard and misunderstood; they treated her qua trans person in a category of no import; and they seemed to view her testimony as not credible and her character as manipulative and disingenuous. Mechanisms of silence have effects on each of these domains, as well as economically. These domains overlap and shift in impact. Note that the rise of neoliberalism is relevant here: knowers are (among other things) people who are independent and can stand on their own; they are not needy, dependent, or chronically suffering, as many of those with mental illnesses are viewed. Since neoliberalism washes its hands of the poor (ostensibly "allowing" them to buy goods from the private sectors), and since mental illness often leaves people vulnerable to homelessness and poverty through job loss, high costs of insurance, and sometimes difficulty in functioning, neoliberalism strikes at the heart of epistemic inequalities in addition to so many other forms in inequality. Here, I focus on epistemic and ethical aspects of silencing and then broaden its features and effects afterward.

As José Medina writes, "In a situation of oppression, epistemic relations are screwed up" (Medina 2013: 27). Kristie Dotson introduced the concept of epistemic violence to explain the phenomena that marginalized and disadvantaged groups experience with respect to communicative interactions. **Epistemic violence** occurs when (a) knowledge that is possessed by subjugated peoples is suppressed or discounted; or (b) when listeners refused to communicatively reciprocate in those interdependent exchanges (Dotson 2011).

One form of epistemic violence that Dotson discusses is testimonial quieting. **Testimonial quieting** occurs when the speaker's credibility as a knower is undervalued due to controlling images. I have argued elsewhere that Black girls are disproportionately silenced in this way in the classroom (Potter 2016). Black girls frequently are perceived as loud and aggressive in schools. Some students may, indeed, be expressing culturally inflected ways of communicating, and some may be exhibiting a politically informed way of surviving as minority students, but the reality is that controlling images seem to dominate teachers' ways of relating to Black girls in the classroom. Research suggests that many teachers try to inculcate a white, middle-class norm of femininity that is docile and quiet. They are more concerned with teaching Black girls social skills—such as not being "loud"—and worry less about their academic development. They tend to see Black girls as outspoken and confrontational, a challenge to teachers' authority. For instance, one teacher scolded a Black girl for calling out the answer to a problem. Other teachers viewed Black girls as disruptive of the classroom and one teacher described them as "very defiant" (Morris 2007). This is an example of testimonial quieting and, as Black girls try to acquire an academic education, such experiences accumulate in forms of epistemic violence.

We see this silencing in mental health care, as well. FTM writer Jacob Hale describes the pathologizing that occurs to obtain medical body enhancement technologies: transpersons are required to receive a *DSM* diagnosis prior to any surgical or hormonal interventions. Hale notes the irony in having to submit to such humiliations to exercise agency over one's own body and, in doing so, in being complicit in one's own erasure. The stigma of mental illness follows one around, rendering one's voice the mere rantings of a crazy person (Hale 2009: 47). Eleanor, in emergency psychiatry, is far from trying to obtain medical interventions—she wants to be safe from harm, to find a refuge for the night. Eleanor does not want to be understood as having a psychiatric illness, as many trans people are; instead, she seems to be crying out not only to be safe but also to be understood from the perspective of *her* experience, not from the perspective

of either a diagnostic gaze or from that of a suspicious health care team that views some visitors to emergency psychiatry as trying to game the system. Yet, she, too, experiences herself as silenced and erased.

Epistemic violence need not be intentional. It arises out of what Dotson calls "pernicious ignorance" which involves "a maladjusted sensitivity to the truth with respect to some domain of knowledge" (Dotson 2011: 241–242). It is a kind of *situated* ignorance, meaning that it follows from one's social position and/or epistemic location and fosters significant epistemic differences among diverse groups. Epistemic difference is the gap between different worldviews caused by differing social situations (economic, sexual, cultural, etc.) that produce differing understandings of the world, differing knowledges of reality. Eleanor, for example, knows what it is like to be a homeless trans person in fear of bullying and attack from others, while the trained and well-intended health care team failed to grasp what life would be like for her or others in similar circumstances, largely because their knowledge-base is so different. What Eleanor knows about the local world and what the health care team know about their patients seems an unbridgeable gap. In other words, epistemic violence is partly explained by the social location of the advantaged and privileged individual or group, where the listener's failure to give uptake to the speaker is done out of ignorance when one could have, and should have, known better. Hence, the ignorance is pernicious even when it is unintentional.

This point is important because it highlights the connection between epistemic and ethical responsibility. Epistemic violence is a form of oppression, and it is fueled by the very sorts of stereotypes and controlling images I discussed in the section Mental Illness and Mechanisms of Oppression and is made manifest in our cognitive, affective, and embodied ways of perceiving and relating to others. Pernicious ignorance is bad both because it is epistemically irresponsible and because it is morally wrong, because it occurs as a pattern of violence. It is violent and, thus, harmful because it harms members of a group's ability to speak and be heard. More specifically, it (1) damages her confidence in her beliefs and her confidence in herself as a knower; (2) undermines her sense of herself as an epistemic agent and, hence, her humanity; (3) hinders the development of intellectual courage; and (4) "excludes the subject from trustful conversation" (Dotson 2011; Fricker 2007: 54).

I suggest that, if we are to engage in changes in clinical encounters in ways that address stereotyping, microaggressions, and epistemic violence, we who occupy social locations of privilege and advantage will need to (among other things) do "world"-traveling: we will need to learn to de-center ourselves, our

ways of perceiving the world, our ease and comfort, and our confidence in our own rightness. We will need to learn to see ourselves the way that those who are disenfranchised and systematically discriminated against see us. We will need to learn to be critically reflective and to interrogate assumptions about the norms that govern social, legal, economic, and political structures.

While this is a process that is relevant to all of us, it is especially important work for the privileged to do. Medina's work suggests that this process of change may be substantial and difficult. "Active ignorance has deep psychological and sociopolitical roots: it is supported by psychological structures and social arrangements that prevent subjects from correcting misconceptions and acquiring knowledge because they would have to change so much of themselves and their communities before they can start seeing things differently" (Medina 2013: 58). Specifically as it pertains to listening to those who are experiencing abandonment, silencing, and deep distress, it may take significant moral effort. As Jill Stauffer writes, "Only a self capable of being jolted out of its mundane complacency is up to the task of both hearing what repair demands and helping to invent new responses to harms that no preexisting remedy fully comprehends" (Stauffer 2015: 7).

Hearing the cries of others' pain and suffering, and creating openings for others even to make those cries, are constitutive elements of psychiatry. But are the moral demands I've described in this chapter truly relevant to clinicians? In the last section, I will locate some of these issues within the geography of psychiatry and the work of helping those with mental distress and disability navigate to less isolated and lonely places.

Application to psychiatry

Patients can be silenced and left struggling with their illness, their positionalities within larger society, and the epistemic and ethical harms they may experience. This happens even when psychiatrists are well-intended, empathetic, and engaged in efforts to relieve patients' suffering. Systems of oppression and domination create *patterns* that are found in different social groupings, including not only of the privileged but also within professional roles. The social grouping under consideration in this section is that of clinicians. The social conditions I have discussed in this chapter can affect our capacity to hear and to be heard correctly, capacities that are of central importance to good practices in psychiatry and other mental health practices. The social position of privileged people, including

that of psychiatrists, affords them the ability not to know certain things and the assumption that they do not need to know.

As I have indicated, structural violence and intersecting oppressions, in tandem with neoliberalism, shape the experiences one has, the kinds of concepts one forms, and the beliefs one holds about what is true about the world. Such experiences are formed in patterns of differentially situated people, wherein the locations one occupies have real-world material, epistemic, ethical, sociopolitical, and economic consequences. In psychiatry, one effect is that of mechanisms of silencing. Eleanor, in the opening case, *suffers* silencing, further alienation, and the very real threat to her personal safety. Historically, the phenomenon of silencing can be seen when the threat to white people that Black power became during the Civil Rights Movement prompted a new diagnosis called "the protest psychosis." This diagnosis gave rise to an increase in the number of African American males viewed as schizophrenic (Metzl 2009). Jonathan Metzl argues that the increase in diagnoses among this group was political, not clinical (even if many clinicians believed they had identified a new clinical phenomenon), but it functioned as an explanation for why Black people were protesting against white racism: they were paranoid, suffering delusions of persecution. Protesting against racism became dangerous and ill-advised for African Americans because they might end up diagnosed, involuntarily hospitalized, and medicated for an illness many of them did not have.

Socially and professionally developed tendencies to participate in epistemic ignorance are a vulnerability for psychiatrists and other mental health professionals because their training in primarily biological psychiatry, an emphasis on Evidence-Based Medicine as the gold standard, and the increasing influence of neoliberal medical corporatism often work together to create clinicians who are expected—indeed, may even be required—to narrowly focus on the person as a generic individual as they decide whether to diagnose or not and to treat efficiently enough so as to maximize remuneration. Because hospitals and clinics increasingly are driven by profit-making, most psychiatrists simply do not have sufficient time to devote to each patient. Those clinicians who reject these constraints may go into private practice—but this may preclude their taking insurance and, hence, reinforces a climate of care for the wealthier and neglect of the poor. This focus may reinforce some clinicians' largely unconscious habits of stereotyping because the use of stereotypes is quicker and more time-efficient. These practices, then, constrict many clinicians' access to accurate and complete knowing. Thus, clinicians may elide or silence many crucial factors that influence the experiences and needs of the person in front of them.

As this chapter explains, structural violence, intersecting oppressions, and the rise of neoliberalism make it more likely for clinicians not to understand how the professional norms to which we subscribe may be inadvertently harming others. They make it difficult to see how and when we act with epistemic violence and other, sometimes subtle acts of microinvalidations. As Stauffer argues, what those suffering from alienation and experiences of abandonment need is help with *world-building*—the broad socioeconomic as well as minute work of helping people put their lives back together, or to help them manage them to the best of their ability—and this takes "broad social support that functions as a promise that, though she was once abandoned by humanity, that will not be allowed to happen again" (Stauffer 2015: 7).

Following Lugones, I advocate the willful flexibility of shifting from one construction of life, where one is more or less comfortable, to an other construction of life, where one is less at ease. This dis-ease in the clinician arises from the experience of decentering oneself enough to be able to grasp how the person with whom one is interacting has been, and currently is, experiencing her "world/s." The way to travel is animatedly and lovingly. (I say more about this idea in Potter 2003.) The clinician needs to remain oriented to oneself, but flexible and conscientious enough that one is able to hear and perceive the other's communications without imposing one's own lexicon and without automatically drawing down the instituted social imaginary. The space between clinician and the person/patient is an intersubjective space, "freighted with a giving of signs that precedes communication in language, as well as the possibility of bringing 'me' out of myself and into a world" (Stauffer 2015: 25). Stauffer emphasizes that good listening also involves attending not only to what is communicated by the other but to what is not being heard—by both parties. I would add that, in the case of power differentials such as found between clinician and person/patient, it especially is important for clinicians to be attentive to what they may not be hearing—a difficult and paradoxical task indeed. To do clinical work well, then—meaning that engagement with others alleviates suffering and may provide healing and treatment that others need—also requires that clinicians understand how oppressions work and how their own social locations may reproduce, or undermine, their force within psychiatry and psychology. These are epistemic and ethical responsibilities of clinicians.

This is a lot to ask of clinicians, and I am not hopeful about the world's ability to undercut the force of neoliberalism. To the extent that health care is part of corporatization and privatization, it's unclear that democracy on the ground can gain traction sufficient to provide liberatory relief from human suffering. But to

the extent that it might provide positive shifts, it will take psychiatrists working together to "world"-travel, and to imagine and create, together with the people who they are there to care for and to provide relief from ethical loneliness, new terrain.

Notes

1 On the responsibility being more on the shoulders of the relatively privileged and powerful, see Lugones and Spelman (1983) and Potter (2002).

2 Between the years 1976 and 1983, tens of thousands of people in Argentina were kidnapped, tortured, and murdered. Some were held in cells, while others were drugged and dropped from airplanes over the ocean. Many victims' remains have never been found.

References

Brown, W. (2015), *Undoing the Demos: Neoliberalism's Stealth Revolution*, Brooklyn, NY: Zone Books.

Burns, P. and Gimpel, J. (2000), Economic Insecurity, Prejudicial Stereotypes, and Public Opinion on Immigration Policy, *Political Science Quarterly*, 115:201–225.

Carastathis, A. (2014), The Concept of Intersectionality in Feminist Theory, *Philosophy Compass*, 9:304–314.

Charlton, J. (2000), *Nothing about Us without Us: Disability Oppression and Empowerment*, Berkeley and Los Angeles: University of California Press.

Collins, P. H. (2000), *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*, 2nd edition. London: Routledge.

Corrigan, P. (2004), How Stigma Interferes with Mental Health, *American Psychologist*, 59:614–625.

Crenshaw, K. W. (1989), Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics, *University of Chicago Legal Forum*, 140:139–167.

Crenshaw, K. W. (1991), Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color, *Stanford Law Review*, 43:1241–1299.

Cudd, A. (2006), *Analyzing Oppression*, Oxford: Oxford University Press.

Dotson, K. (2011), Tracking Epistemic Violence, Tracking Practices of Silencing, *Hypatia*, 26(2):326–257.

Farmer, P., Nizeye, B., Stulac, S., and Keshavjee, S. (2006), Structural Violence and Clinical Medicine, *PLoSOne Medicine*, 3:e449.

Fricker, M. (2007), *Epistemic Injustice: Power and the Ethics of Knowing*, New York: Oxford University Press.

Garry, A. (2012), Who Is Included? Intersectionality, Metaphors, and the Multiplicity of Gender, pp. 492–530. In *Out from the Shadows: Analytical Feminist Contributions to Traditional Philosophy*. S. L. Crasnow and A. M. Superson, eds. New York: Oxford University Press.

Gilman, S. (1985), *Difference and Pathology: Stereotypes of Sexuality, Race, and Madness*, Ithaca, NY: Cornell University Press.

Grasswick, H. (2012), Knowing Moral Agents: Epistemic Dependence and the Moral Realm, pp. 307–338. In *Out from The Shadows: Analytical Feminist Contributions to Traditional Philosophy*. S. L. Crasnow and A. M. Superson, eds. Oxford: Oxford University Press.

Hale, J. (2009), Tracing a Ghostly Memory in My Throat: Reflections on FTM Feminist Voice and Agency, pp. 43–65. In "*You've Changed": Sex Reassignment and Personal Identity*. L. Shrage, ed. Oxford: Oxford University Press.

Klein, N. (2007), The Shock Doctrine: The Rise of Disaster Capitalism, New York: Picador Press.

Kukla, R. (2014), Performative Force, Convention, and Discursive Injustice, *Hypatia*, 29:440–457.

Langton, R. (1993), Speech Acts and Unspeakable Acts, *Philosophy and Public Affairs*, 22:293–330.

Lugones, M. (1987), Playfulness, "World"-Traveling, and Loving Perception, *Hypatia*, 2:3–19.

Lugones, M. and Spelman, E. (1983), Have We Got a Theory for You! Feminist Theory, Cultural Imperialism and the Demand for "The Woman's Voice," *Women's Studies International Forum*, 6:573–581.

McGregor, S. (2011), Neoliberalism and Health Care, *International Journal of Consumer Studies*, 25:82–89.

Medina, J. (2013), *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and the Social Imagination*, Oxford: Oxford University Press.

Metzl, J. (2009), *The Protest Psychosis: How Schizophrenia Became a Black Disease*, Boston, MA: Beacon Press.

Morris, E. (2007), "Ladies" or "Loudies"? Perceptions and Experiences of Black girls in Classrooms, *Youth & Society*, 38:490–515.

Mouffe, C. (1996), Democracy, Power, and the "Political," pp. 245–256. In Democracy and Difference: Contesting the Boundaries of the Political. S. Benhabib, ed. Princeton, NJ: Princeton University Press.

O'Neill, S. (2010), Struggles against Injustice: Contemporary Critical Theory and Violence, *Journal of Global Ethics*, 6:127–139.

Potter, N. N. (2000), Giving Uptake, *Social Theory and Practice*, 26:479–508.

Potter, N. N. (2002), *How Can I Be Trusted? A Virtue Theory of Trustworthiness*, Lanham, MD: Rowman-Littlefield.

Potter, N. N. (2003), Moral Tourists and World-Travelers: Some Epistemological Issues in Understanding Patients' Worlds, *Philosophy, Psychiatry, and Psychology*, 10:209–223.

Potter N. N. (2015), Reflections on Things We Don't Want to Think About: Intersection of Colonialism, Transgenerational Trauma, and Oppression within Psychiatry, *Journal of Ethics in Mental Health*. https://www.jemh.ca/issues/v9/documents/JEMH_Open-Volume_Article_Theme_Colonization_Reflections_June2015.pdf.

Potter, N. N. (2016), *The Virtue of Defiance and Psychiatric Engagement*, Oxford: Oxford University Press.

Sen, A. (2009), *The Idea of Justice*, Cambridge, MA: Harvard University Press.

Stauffer, J. (2015), *Ethical Loneliness: The Injustice of Not Being Heard*, New York: Columbia University Press.

Sue, D. W., Cpodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A. M., Nadal, K. L., and Esquilin, M. E. (2007), Racial Microaggressions in Everyday Life: Implications for Counseling, *The American Psychologist*, 62:271–286.

Tessman, L. (2009), Feminist Eudaimonism: Eudaimonism as Non-Ideal Theory, pp. 47–58. In *Feminist Ethics and Social and Political Philosophy: Theorizing the Non-Ideal*. L. Tessman, ed. Dordrecht and New York: Springer.

16

Mental Illness, Agency, and Responsibility

Michelle Ciurria

Introduction

There is a widespread assumption among philosophers that people with mental illness are not responsible and not apt targets of responsibility attributions such as praise and blame (e.g., McKenna 2012; Strawson 1963; Watson 2013). Maybe these theorists intend their comments to apply only to the most debilitating cases of mental disorders, but these are not the *typical* cases, and emphasizing them can lead to the misconception that people with mental disorders have no responsible agency at all. In fact, people with mental disorders are a very diverse population with diverse capacities, and most are capable of exercising responsibility over a substantive range of choices. This is recognized by the leading contemporary approach to mental health, that is, the client-centered approach popularized by Carl Rogers (1951, 1995), which treats mental-health service users as presumptively capable of "responsibility and self-advocacy" (Mental Health Commission of Canada 2015).

Even if some types of mental illness are incapacitating, this depends on the type of disorder and the individual's personality: the current version of the *Diagnostic and Statistical Manual* (*DSM*-5) includes over 250 different mental disorders, with various symptoms, and all of the disorders are responsive to some treatment modalities. Moreover, treatment-responsiveness depends not only on the type of disorder, but also on the individual's personality and circumstances. It is therefore not plausible to think that everyone with a mental disorder is non-responsible, *full stop*. As I argue in this chapter, when discussing responsibility and mental disorders we should avoid a "blanket exclusion clause"[1] that exempts all mentally ill people from the reciprocal relationships whereby we hold one another responsible.

Philosophers who write about responsibility typically focus on *moral responsibility,* which is defined as the capacity to respond to moral considerations (see especially Fischer 2006, 2011). But the operative notion of responsibility can be taken to be broader than this, encompassing responsiveness to *any* considerations that enhance agential functioning. People can be held responsible *simpliciter* for responding or failing to respond to non-moral reasons. For example, we might say that Van Gogh was *responsible simpliciter* for creating innovative artworks, though this is not a moral type of responsibility. And we might also be inclined to say that he was praiseworthy for his artistic accomplishments in spite of the fact that he suffered from severe schizophrenia (Blumer 2002). In what follows, I will use the term "responsible" in this broad sense, to cover both moral and non-moral exercises of agency, and I will argue that service users can be held responsible in both of these senses.

In this chapter, I survey the dominant theories of responsibility and select the candidate that is best able to resist the blanket exclusion clause. I argue that the best candidate is a version of Manuel Vargas's agency-enhancement view (AEV), on which responsibility attributions aim to enhance agency in members of a particular community. AEV justifies responsibility attributions (praise, blame) on the basis of whether they are likely to enhance the target agent's functioning, regardless of the person's mental-health status; thus, it cuts across diagnostic categories and the normative-versus-disordered divide. This view is also appealing because it allows us to see service users (i.e., clients of mental-health systems) as amenable to responsibility attributions in therapy so long as these attributions enhance their functioning, consistent with the client-centered view. This expands the scope of responsibility to include service users both in the community and in mental-health systems.

After defending AEV, I consider whether it can be used to redeem Strawson's view from the blanket exclusion clause, and the related idea (proposed by Strawson) that the primary purpose of treatment is to exclude people from the moral community. (Strawson famously argues that mentally ill people are not moral agents and not apt targets of praise and blame.) This view stands in opposition to the client-centered therapeutic model, which sees therapy not as a means of excluding and isolating service users, but rather as a means of *re-integrating* them into their local communities by empowering them to pursue meaningful social and occupational projects. I argue that on a synthesized Vargas-Strawsonian view, we can redefine *therapy* as internal to our shared practice of holding people responsible, as opposed to a mechanism for excluding service users from this network of reciprocal relationships.

In developing a Vargas-Strawsonian synthesis, I make two distinct claims about the relationship between responsibility and therapy: (1) the reactive attitudes of praise, blame, and the like are fitting in therapeutic contexts insofar as they enhance the service user's agency—in other words, service users can be held responsible for their choices and actions; and (2) therapy is itself a distinctive type of reactive attitude with sui generis properties that distinguish it from other types of reactive attitude. These claims allow us to see anyone with a treatment-responsive mental disorder as a responsible agent, capable of exercising control over a significant range of choices.

In the last section of this chapter, I discuss how we can express suitable responsibility attributions toward service users. I argue that such responsibility attributions should be particularist rather than generalist. By "generalist" I mean attributions that flow from *general principles* that apply to all mentally ill people or even entire diagnostic categories, ignoring the particularities of each case. Thus, generalist attributions treat each service user as a generic member of a type of mental illness. By "particularist" I mean a view on which we evaluate responsibility on a case-by-case (particular) basis. I contend that we must take into account two sets of considerations in a particularist assessment: First, *generic considerations* about the type of mental disorder in question and the disorder's clinical response rates, and second, *specific considerations* about the role of the disorder in the client's cognitive economy, the presence of risk factors, and the availability of treatment services in the person's community. The combination of these considerations enables a substantively particularist response to service users, which refuses to paint them all with the same brush.

Definitions

Before beginning, I'll clarify the main terms of my discussion.

I define the terms "mental disorder" and "mental illness" (used interchangeably) according to *DSM*-5 criteria, with the qualification that I am not committed to the "biomedical model," which depicts mental disorders as reducible to physical processes (e.g., pathology, biochemistry, physiology). Fortunately, *DSM*-5 is not committed to this model either. It defines a mental disorder as a "disturbance" in a person's "cognition, emotion regulation, or behaviour," which causes "significant distress or disability in social, occupational, or other important activities," where these experiences depart significantly from cultural expectations (20). (The cultural exemption allows us to exclude normal

grief and other culturally normative experiences.) This definition is found under the "criterion of clinical significance," which captures the *normative* dimension of mental illness: what makes a cluster of symptoms a *dysfunction* as opposed to a normative trait. Most of the rest of the manual is devoted to *diagnostic* criteria, which lay out the characteristic symptoms of various disorders—their "symptomatic profile." Both criteria must be met for a mental state to count as a mental disorder. *DSM*-5 admits that an adequate client case formulation and responsible treatment plan cannot be arrived at on the basis of biological markers alone; rather, environmental risk and prognostic factors as well as cultural and historical factors must be taken into account. This comprehensive reading makes *DSM*-5 look more like a biopsychosocial model than a single-factor biomedical model, and I will assume this multifactorial interpretation of mental illness here.

I will use the terms "clients" and "service users" to describe anyone who uses *mental-health systems*, which encompasses mental-health hospitals, treatment programs, and therapy.

For simplicity, I will sometimes refer to the agency-enhancement view of responsibility as "AEV" and the objective attitude as "OA."

Finally, I will use "responsibility attributions" and "reactive attitudes" interchangeably to denote attitudes of moral and non-moral blame and praise, approbation, resentment, admiration, disapprobation, and related evaluative stances. (I discuss the make-up of the reactive attitudes more specifically below).

Responsibility and the blanket exclusion clause: P. F. Strawson

Many responsibility theorists suppose that, or speak as if, mental illness negates responsibility. P. F. Strawson (1963), for example, delineates two moral stances that we can take toward other human beings: (1) the "participatory stance," in which we deploy the reactive attitudes toward fellow community-members with whom we can establish relationships of reciprocal trust and mutual obligation, and (2) OA, which we are supposed to take toward non-responsible agents. The second category includes the "psychologically abnormal" and "compulsives," in Strawson's words. (I will use theorist's terms whether I endorse them or not, noting that some of these terms are anachronistic and insensitive.) When we assume OA, we treat the target agent as an "object of social policy …, management, treatment, and control," and as an outsider to the "web of

attitudes and feelings" that bind people together in community (Strawson 1963: 13). That is, OA functions to *exclude* people from the community by denying them in-group-member status. The participatory stance is appropriate for actual or potential in-group members only. Strawson places mentally ill people in the "objective" realm, which deploys treatment, management, and control as mechanisms of exclusion.

Watson and McKenna build on Strawson's theory, suggesting certain revisions but preserving the core idea that "exempting conditions" apply to "children," the "psychotic," and "sociopath[s]" (Watson 2013: 119), as well as "the severely mentally retarded (sic), the insane, those suffering from pervasive delusions, and so on" (McKenna 2012: 76). This seems to suggest that these groups all belong to the objective category, regardless of differences between different categories of mental illness and differences between individuals within each category. People who adopt a Strawsonian model inherit from Strawson the burden of explaining why people with mental disorders, particularly clinically significant mental disorders that impair the person's relational competency and emotional wellbeing, do not collectively fall within the remit of the objective attitude. Our experiences may conflict with the notion that these individuals are outside the scope of our gratitude, approbation, resentment, and disapprobation. My aim is to provide philosophers with resources to avoid this blanket exclusion, by showing that therapeutic engagement is internal to the participatory stance.

In the next section, I survey influential accounts of responsibility in search of a view that avoids distorting and homogenizing claims about mental illness. I consider three candidates: the character view, the control view, and the AEV.

Character, control, and agency enhancement

Character theory

Character theory is a very influential approach with roots in Harry Frankfurt (1969). Modern character theorists hold that people are responsible for actions and omissions that issue from their character (e.g., Arpaly 2015; Sher 2009; Smith 2005).

While there are differences among these theorists, they all tie responsibility to features of a person's deep self—a core cluster of integrated neurophysiological states. For example, Sher holds that we are only responsible for actions that stem from a causally integrated set of "psychological or physiological states" that are

"among the constitutive features of a particular self"(2009: 121). This suggests that a person's "responsible self" is an interconnected web of neurophysiological states that enable relatively stable patterns of thought and action—patterns that allow us to identify the person as a unique individual.[2]

The criterion of stable character seems to imply that, of all the categories of mental illness, only people with *personality disorders* are responsible for their choices, since a personality disorder is defined as "an enduring pattern of inner experience and behaviour that … is pervasive and inflexible …, [and] is stable over time" (APA 2013: 646). By comparison, non-personality-type disorders are relatively episodic, cyclical, and situation-sensitive; bipolar disorder, for example, involves cycles of manic, hypomanic, and depressive episodes. Personality disorders, in fact, can be more stable than many normative personality traits (e.g., introversion), especially if left untreated. But people with personality disorders are hardly paradigm cases of responsibility! Responsibility is generally assumed to be something worth wanting, whereas personality disorders—which cause dysfunction and distress by definition according to *DSM*-5—are not particularly desirable.

Giving personality disorders special status while completely exempting people with cyclical and episodic disorders, moreover, bizarrely treats the first population as emblematically responsible while treating the second population as utterly non-responsible. This dichotomy flies in the face of the general consensus that disorders lie on a spectrum, rather than forming a neat binary dividing functional from non-functional people.

These considerations give us good reason to canvass other approaches.

Control theory

The dominant competing alternative to character theory is the "control view," which was popularized by John Martin Fischer (2006, 2011). Fischer holds that we are responsible for actions that are under our control at the time of action, or were amenable to control prior to that time. This view is "'historical' in a strong way" in the sense that it allows us to trace responsibility back to prior moments of choice (Fischer 2006: 236). Hence, someone might be responsible for committing an indiscretion while drunk if he could have avoided becoming intoxicated in the first place, but he is excused if someone surreptitiously spiked his drink at a party.

This theory directs our attention to considerations about whether someone has control over her actions and their neurophysiological substrates. Although Fischer (like Sher) makes few direct references to mental disorders, his theory implies that responsibility for the effects of a mental disorder ultimately depends

on whether the relevant neurophysiological states were amenable to control at the time of action or prior to that time.

This criterion is appealing insofar as it cuts across diagnostic and clinical categories, since we can evaluate neurophysiological states on a control basis irrespective of their mental-health status. (For example, whether someone with schizophrenia has control over her choices depends on her specific psychological profile, not the symptomatic profile of schizophrenia considered in the abstract.) Control, then, is a prima facie good candidate for a particularist (case-sensitive) approach to mental illness.

The main drawback of the control view is that it is, on scrutiny, *too historical.* When assessing whether someone is responsible on the control view, we are only meant to countenance the person's states *at the time of action and prior to that time.* This might not be the best approach given that many non-personality-type disorders are highly variable across time and contexts. For example, bipolar disorder features cycles of manic, hypomanic, and depressive episodes; depression can occur after a pregnancy (postpartum) or seasonally (seasonal affective disorder); anxiety can be triggered by stressful life events and personal loss. And so on.

This raises some interesting questions. If someone commits a problematic action during a manic episode, should we hold the person responsible on grounds that she could have exercised more control over her choices at the time of action or prior to it? What if the disorder has since gone into remission, and we are in the position of making a post facto attribution of responsibility? Post facto criticism for a disorder that *has since gone into remission* hardly seems helpful, and it may actually be harmful if the person is highly sensitive to blame. (Notably, some disorders cause excessive and irrational self-blame—for example, posttraumatic stress disorder and persistent complex bereavement disorder. Blaming people with these disorders would be counter-therapeutic.) Alternatively, we could exempt everyone with a cyclical/episodic disorder on grounds that these populations lack any degree of control whatsoever, but then we are forced back onto a blanket exemption for these groups.

Thus, it is worth exploring whether there is a more suitable theory of responsibility—preferably one that neither exempts entire populations, nor blames people for the effects of conditions that are currently being well managed.

It may be instructive here to look at therapeutic models of responsibility, especially if we are looking for a theory of responsibility that is compatible with therapy. Modern therapy is essentially outcome based: it seeks to facilitate positive treatment outcomes on the basis of evidence. This is why *DSM*-5 includes the

criterion of clinical significance, which draws attention to the clinician's guiding aim of restoring or enabling a minimal threshold of social and occupational functioning.

This aim does not fit well with a *historical* approach to responsibility, which does not prioritize outcomes, but instead limits responsibly relevant considerations to conditions at the time of, or prior to, the action under consideration. Contrary to this backward-looking orientation, when clinicians make a clinical case formulation, they take into account the client's relevant history *in conjunction with* a range of non-historical factors, including "presenting concerns" (the client's state of mind and therapeutic goals at the time of clinical presentation), present cultural factors, social systems, possible intervention strategies, and *any other factors* that might help the therapist promote the client's well-being (APA 2013: 19). These considerations clearly *outstrip* past and present factors. A temporally restricted theory of responsibility, therefore, is incompatible with the therapist's professional commitments.

The most suitable model for the contemporary practicing therapist must be an outcome-based one.

The agency-enhancement view

To recap: we are looking for a model of responsibility that is admissible in therapeutic contexts—one that does not deem current service users categorically non-responsible, and one that respects the outcome-based orientation of modern therapy. A good candidate is Vargas's AEV.[3]

On Vargas's view, responsibility attributions are fitting if they are the type of attitude that is likely to enhance agency in members of the community. The advantage of this view for our purposes is that it is outcome based, in the sense that it seeks to enhance "valuable form[s] of agency" (2013: 173), and this is consistent with the aims of contemporary therapeutic models. Interestingly, Vargas's view can be seen as construing all fitting responsibility attributions as *essentially therapeutic*, inasmuch as they must function to enhance the agency of the target individual—otherwise they are unjustified and inadmissible.

This "functionalist" construal of the reactive attitudes collapses the responsibility-therapy divide posited by Strawson, not simply by integrating praise and blame into therapeutic contexts, but by defining praise and blame as *essentially therapeutic*. Since responsibility attributions have the same fittingness conditions and primary function as therapy (i.e., enhancing agency), it follows that responsibility attributions are admissible in therapy. Vargas does not

explicitly say that praise and blame are therapeutic attitudes, but I take this to be a plausible reading of his view, and one that accommodates the particularist approach to mental illness that we are after.[4]

The idea that therapy can involve praise and blame runs counter to the long-standing philosophical orthodoxy, which is (perhaps) nowhere as apparent as in Strawson's view that "treatment" is a form of OA, alongside techniques of "management" and "control." The reactive attitudes of praise and blame, by contrast, operate in the sphere of the "participatory stance," which solidifies our social bonds.

For Strawson, OA functions to treat, manage, and control people who are insensitive to the norms and practices of the general community, but OA is also described by Strawson as a mechanism for *excluding* non-responsive individuals from the community. That is, OA serves the dual function of excluding people on the one hand, and managing, controlling, and treating people on the other. These functions are not easily reconciled if we think of therapy in client-centered terms, as intended to help the agent pursue an occupation and cultivate meaningful relationships.

The modern approach to therapy is therefore diametrically opposed to Strawson's envisioned purpose of "treatment" as a mechanism for controlling, managing, and isolating people to protect the general community. While it is not hard to see why violent criminals should be detained and managed for the sake of the community, this treatment is not appropriate for the vast majority of mental-health service users. We therefore need to dissociate mechanisms of treatment from mechanisms of management and control, since the two categories do not serve the same purposes in our society. (That said, therapy can be integrated into prison systems.)

Why would Strawson think of therapy as an OA? It is possible that he had in mind Michel Foucault's notion of "therapy" as practiced in the asylums of the pre-modern era, where asylums functioned as mechanisms of discipline and punishment, designed to control and manage the mentally ill along with other stigmatized social groups. But these practices are now recognized as human rights violations and barred from use. Modern therapy, in stark contrast to pre-modern asylums, ideally seeks to respect people's dignity and autonomy.

It is also possible that Strawson only had in mind *passive* modes of treatment, such as pharmaceutical drugs that operate passively on the client's brain without her conscious intervention; but this is just *one* type of treatment modality, and it is supposed to be delivered in the context of a dynamic relationship between

therapist and client. Modern treatment principles recognize that, if an agent is decisionally capable, *active cooperation and consent* are needed for any treatment modality to be effective.

This means that therapy as practiced in the modern era should fall within the participatory stance, not the OA, since modern therapy aims to enhance agency and integrate clients into their communities, not to exclude and "manage" them.

We can revise Strawson's conception of the reactive attitudes to fit modern principles of therapy by giving these attitudes a Vargas-style justificatory basis, that is, by reimagining them as justified by their propensity to enhance agency. On this *hybrid view*, the reactive attitudes are fitting responses to treatment-responsive service users, that is, almost all service users. (Decisionally incapable people may be a borderline case, as they can be subjected to involuntary treatment under the law, though this population could still retain control over a non-trivial range of choices.)

In the next section, I discuss how the hybrid approach allows us to tailor the reactive attitudes to a person's unique psychological profile, rejecting the one-size-fits-all approach that is necessitated by some other views (particularly character theory).

A functionalist strawsonianism

How does AEV fit with Strawson's picture of the reactive attitudes?

The paradigmatic reactive attitudes described by Strawson (1963) in his landmark paper are "gratitude, resentment, forgiveness, love, and hurt feelings," but the emphasis of the paper, entitled "Freedom and Resentment," is clearly resentment and the other "negative" reactive attitudes. Indeed, the negative reactive attitudes have been the focus for most philosophers, who tend to marginalize the positive reactive attitudes. For example, Susan Wolf (2011) focuses on anger, and Macalaster Bell (2013a, b) emphasizes hostile attitudes and contempt. Some, like Sher (2009), think that praise and blame have different warranting conditions, contrary to Strawson's integrative picture.

It is not hard to see, however, how negative *and positive* attitudes could fit together in a comprehensive set of "social practices" (Strawson 1963)—or a shared "responsibility system" (Vargas 2013: 6)—that functions to enhance the agency of its members through the operation of different attitudinal valences. Praise may be fitting for people with deficits in praise-sensitivity (e.g., low self-regard), while blame may be fitting for people with deficits in blame-sensitivity (e.g., narcissism). More precisely, praise and blame may be two sides of the

same coin, or two parts of the same system, working together to facilitate an equilibrium of positive and negative self-regarding attitudes in target agent, correcting for psychological imbalances.

If responsibility attributions were exclusively negative, or even primarily negative, one might be tempted to think that the reactive attitudes are "never appropriate during therapy," as Mike W. Martin says of moral praise and blame, along with all "explicitly moral language" (2012: 39). Martin is right that blame might be harmful for many service users, given that mental illness is widely stigmatized and blame might exacerbate the effects of this stigmatization. But even if we grant that therapists ought to suspend *blame* toward service users, it is hard to find a rationale for *also suspending praise*. If praise is admissible in therapy, then reactive attitudes (which include praise) are thereby admissible.

Surely the best approach is a nuanced one that responds to the service user's particular mental states. Some mental disorders characteristically involve self-blaming symptoms: for example, *persistent complex bereavement disorder* involves "maladaptive appraisals of oneself in relation to the deceased or death, for example self-blame" (APA 2013: 790). Blame would not be therapeutic for this symptom, though praise might be remedial. Other disorders involve grandiose self-praise: for example, people with narcissistic personality disorder "usually want praise for their 'superiority'" (APA 2013: 669). Praise would not be therapeutic for this symptom, but blame might be useful. In light of the symptomatic profiles of various mental disorders, the idea that praise and blame have *no therapeutic value* is simply not credible.

So far, these considerations show that blame and praise can be appropriate in therapeutic contexts, depending on the client's specific symptoms. This first proposal is relatively modest, and would likely be accepted by many contemporary philosophers. But I also want to make a stronger claim: that "therapeutic engagement"—a therapist's ways of engaging with a client—can constitute a distinctive type of reactive attitude, over and above those cited by Strawson.

The idea that we can expand Strawson's original catalogue of reactive attitudes is not new. Victoria McGeer (2008) describes *substantial trust* as a type of reactive attitude, though it does not appear in any of Strawson's essays. Substantial trust functions to enhance the trustee's self-efficacy, and it is justified even in the face of evidence that the trustee may not be capable of living up to the trustor's level of confidence. That is, substantial trust is "irrational" in the sense that it disregards evidence concerning the trustee's capabilities, but it is *justified nonetheless* by its propensity to increase the trustee's self-regard and self-efficacy.

The trustor, in a sense, takes a leap of faith in investing in the trustee. Substantial trust is not justified by rational appraisal, but rather by the *hope* that personal investment will empower the trustee.

Substantial trust is similar to *therapeutic engagement* in that it is an attitude of supportive investment in another person, which is meant to enhance that person's agency. Substantial trust is not *identical or reducible* to therapeutic engagement, however, because it has unique features—unique emotional, cognitive, and semantic contents. A trustor might invest in someone by, for example, sponsoring her for a job, helping her prepare a resume, or providing helpful career advice. This is not what a therapist does, nor what a generic "praiser" or "approver" does; substantial trust is distinct from these other reactive attitudes, and is indeed sui generis in its specific contents. Yet it is a species of reactive attitude because it functions to enhance agency.

Likewise, therapeutic engagement is an attitude of engagement that serves to enhance agency in the recipient, but a therapist is not a trustor, a praiser, or an approver. To understand how substantial trust and therapy differ from one another and from the other reactive attitudes, it will be instructive to first clarify what a reactive attitude is. Then we can consider what distinguishes each type of attitude.

Strawson did not clearly define the reactive attitudes, so subsequent scholars have offered their own proposals. There are three main approaches. The first view holds that the reactive attitudes have only "cognitive contents," meaning that they are dispassionate judgments about the target agent's mental states or the fittingness of a particular reaction (Zimmerman 1988; Haji 1998). A second view holds that the reactive attitudes contain both cognitive and emotional contents—a judgment about the target agent together with a corresponding positive or negative feeling (Wallace 1994). A third view holds that reactive attitudes contain only emotional or perceptual contents, not judgments (McGeer 2008). All of these views provide individuating criteria. Rather than selecting among them, I will simply show how *each* of these approaches allows us identify therapeutic engagement as a distinctive type of attitude, allowing us to expand the sphere of the reactive attitudes to therapy.

Allowing that the cognitive and/or emotional contents of the reactive attitudes are what individuate them as distinct states, we can say that praise involves a judgment to the effect that the addressee is praiseworthy, and/or a corresponding positive emotion; blame involves a judgment of censure and/or a corresponding negative emotion; substantial trust involves a judgment to the effect that someone is trustworthy, and/or an attitude of optimism and investment. And so on for the

other attitudes. These attitudes also have distinctive modes of expression. Praise might be expressed in the statement, "Good job," while substantial trust might be expressed in the statement, "I would be happy to mentor you for this job." Thus, these attitudes can be differentiated as distinct types by their cognitive and/or emotional contents, as well as their characteristic modes of expression or semantic contents.

What of therapeutic engagement? The therapist might engage with the client by saying, "you're making good progress in your cognitive behavioral therapy (CBT) workbook," or "why have you stopped coming to group CBT sessions?" These speech acts resemble praise and blame, but they contain *therapeutic references* that are unique to the therapeutic setting. The therapist is in a position to praise the client for therapy-relevant choices and blame the client for therapy-relevant transgressions (and to receive blame and praise for her own choices as therapist). Thus, the attitude of "therapeutic engagement" has unique therapy-relevant cognitive and/or emotional, as well as semantic, contents.

This attitude also has a unique source of justification or *practical authority*, which is the therapeutic alliance between the therapist and the client. A therapeutic alliance is defined by Duff Waring as "a *supportive interpersonal relationship* between a trained therapist and patient (*sic*) in which spoken conversation is used to encourage the communication of conflicts and to develop the patient's insight into the motivation behind her problematic patterns of thought and behaviour" (2016: 1, emphasis mine). Notably, this relationship is an interpersonal one with all the defining features of the participatory stance, and none of the defining features of the OA. The therapeutic alliance grounds reciprocal authority relations between the therapist and the client, which enables the participants to authoritatively hold each other responsible for their choices and actions during therapy. The idea that relationships ground practical authority can be attributed to Darwall (2006), who sees responsibility as agent-relative. Accordingly, therapeutic engagement is a *client-relative attitude*, justified by a reciprocal relationship with the client.

In addition, a therapist might have a special *affective* disposition toward a client, characterized by seeing the client as, above all else, a member of the therapeutic alliance. This permits the therapist to abstract away or ignore the client's therapeutically irrelevant characteristics and choices (as least provisionally), including any moral flaws that are not immediately relevant to the client's treatment plan. But this does not mean that therapy is, as Strawson said, a type of OA: it is simply a unique sort of reactive attitude that prioritizes therapeutic goals.

Therapy still serves to enhance the client's more general agency, including moral agency in many cases—specifically, when a symptom is morally problematic. For example, *intermittent aggression disorder* tends to cause verbal aggression, tirades, fights, outbursts, physical aggression, and damage and destruction of property (APA 2013: 466). When therapeutic interventions treat these symptoms, they also by the same token treat the counter-moral aspects of these symptoms. Thus, while therapeutic engagement can involve a suspension of certain reactive attitudes, it is not incompatible with these attitudes. Specifically, it involves suspension of certain reactions to promote a more fundamental type of agency. Moral reactions may still be appropriate in non-therapeutic contexts and perhaps at a later stage of therapy.

Therapy can be seen as a fitting response—sometimes, perhaps, the *only* efficacious response—to people who have mental disorders that impair their sensitivity to the other reactive attitudes. Rather than treating these individuals as non-responsible, we can see them as responsible for using mental-health resources and exercising control over their treatment plan in order to achieve basic functioning. These choices might empower them to become *more sensitive* to the other reactive attitudes later on.

These remarks show that the therapeutic alliance has unique cognitive and/or emotional and semantic contents, and a special type of practical authority rooted in the therapeutic alliance. Clients and therapists are generally (barring special extenuating circumstances) responsible for the choices they make in the context of this relationship.

In the next section, I close the discussion by exploring how we can express the reactive attitudes toward service users in a suitably particularist (case-sensitive) way—one that respects the individual's unique personality traits and circumstances, rather than reducing the person to a stereotype of "insanity."

Therapy as a reactive attitude: A particularist view

There are two types of consideration that we need to consider when assessing a service user's responsibility: (1) *generic considerations* about the person's type of mental disorder and its treatment-responsiveness, and (2) *specific considerations* about the role of the disorder in the person's psychological profile and its relation to the person's environment. This approach is recommended by *DSM*-5, but it is also the best way to engage with a person as both a service user and a unique human being.

Generic considerations

To begin, when evaluating someone in light of a mental disorder, we should consider the disorder's characteristic symptoms and its general treatment response rates. Treatment-responsiveness is germane because AEV treats responsibility as a matter of responsiveness to the reactive attitudes, including (on my reading) treatment. Thus, if a mental disorder is responsive to a certain treatment modality, the person can be held prima facie responsible for using or failing to use that modality. (This is only a prima facie judgment because we still need to consider individual traits.)

To illustrate the process of generic assessment, let's consider *antisocial personality disorder* (ASPD). ASPD is one of the most treatment-resistant disorders in *DSM*-5, although antipsychotics and mood stabilizers can be used to control the aggression and impulsivity features of ASPD (Coccaro and Kavussi 1997; Walker et al. 2003). In addition, inpatient and residential treatment programs with structured environments and cognitive therapy have yielded some positive results (Davidson and Tyer 1996; Dolan et al. 1997). This suggests that people with ASPD can be held responsible by ordinary community-members for seeking therapy for the treatment-responsive features of their disorder (if treatment is available), and can be held responsible by therapists for their treatment-relevant choices. I say "if treatment is available" because in some jurisdictions health care is not subsidized by the government, which may make treatment inaccessible to low-income groups. In such cases, we must suspend the reactive attitudes and see the person as circumstantially non-responsible. This is not to say that person is *incapable* of acting responsibly, but her responsibility is suppressed by or incapable of being activated due to systems of inequality and injustice.

It should be noted that ASPD is relevantly similar to a non-*DSM*-listed psychological disorder: psychopathy, as defined by Robert Hare's Psychopathy Checklist—Revised (PCL-R). Psychopathy, however, has a higher threshold of antisocial behaviour, including criminal versatility, juvenile delinquency, and revocation of criminal release (Hare 1990). Hare's diagnostic criteria are meant to identify the most violent segment of the prison population, and those who are most likely to rescidivate. People who meet the criteria for psychopathy are shown to be extremely treatment-resistant: they tend to come out of therapy either no different or worse than before (Hare 1999: 198–200; Rice, Harris, and Cormier 1992). They also tend to disrupt treatment sessions and prevent more treatable populations from benefiting (Babiak and Hare 2006: 249;

Hare 1999: 94). Hare therefore recommends that psychopaths not be offered treatment, but instead be incarcerated for as long as possible to protect the public. While psychopaths (fortunately) comprise a very small percentage of human beings—perhaps a low as 1 percent—they make up about 20 percent of the US prison population, and are responsible for the most violent offenses (Hare 1993: 87).[5] This suggests that it might be fitting to adopt OA toward psychopaths as a group because they do not respond to any known treatment and they seem incapable of entering into reciprocal relationships with others. There may be grounds, then, for adopting a *categorical exemption* for psychopaths. Although this goes against strong particularism, it applies to a very small segment of the general population and is warranted on the basis of research on response rates in psychopaths.

Fortunately, all other mental disorders are more responsive to treatment than psychopathy, as well as ASPD. For example, major depressive disorder responds to cognitive behavioral therapy, interpersonal therapy, psychodynamic therapy, work programs, and antidepressant medication in clinical studies, although the efficacy of these modalities depends on the subject's personality: for example, the response rate to CBT in adolescents is higher for those with more rationality and less anxiety and hopelessness (Clarke et al. 1992; Smith et al. 2009). The variable response rates among different sub-populations confirms that therapeutic engagement is fitting for people with depression but must be sensitive to personality traits and the interaction between personality and treatment interventions.

The upshot is that all *DSM*-5-listed mental disorders are at least somewhat treatment-responsive and therefore apt targets of the reactive attitudes, but these attitudes must be sensitive to clinical response rates and the subjective profile of the agent.

Specific considerations

It should be clear by now that generic considerations about a mental disorder's profile and response rates do not settle the question of a service user's responsibility status. We also need to take into account the individual's personality and circumstances.

First, we need to assess the role of a disorder in the person's motivational economy: How strong is its influence over the person? Is the disorder episodic or persistent? Does it have predictable triggers, and can these triggers be avoided or cognitively reappraised? And so on.

Second, we must consider the agent's *external circumstances*. This part of the equation cannot be stressed enough, especially given that it is excluded by the biomedical model. Yet, there is now general consensus that a single-factor explanation of mental illness is empirically flawed and counter-therapeutic. This is why the latest version of *DSM* includes references to environmental risk and prognostic factors, cultural dimensions, and developmental and lifespan considerations. Still, this is a truncated list of environmental factors, which leaves out many fine-grained and underappreciated variables, including some that have only come to light very recently. Researchers have found, for example, that schizophrenia is more prevalent in the UK among the lowest social class in indigenous White peoples and Black immigrant groups (Cooper 2005). This supports "the social causation hypothesis"—the idea that social factors are strongly causally implicated in mental disorders (Cooper 2005: 361). In brief, *low socioeconomic status* (low SES) and *immigrant status* are relevant to explaining mental disorders (although they are not mentioned in *DSM*-5). Since these factors influence reasons-responsiveness in these groups, they are relevant to assessing responsibility.

In the previous section, I said that whether someone is responsible for the symptoms of ASPD depends in part on whether effective treatment modalities are accessible to the person. This is an environmental constraint on responsibility: people are not responsible for treatment-responsive disorders when treatment is not a live option for them. Similarly, environmental risk factors, like low SES, can mitigate or suspend a person's responsibility. The confluence of environmental risk factors together with lack of access to mental-health services may completely defeat a person's control over her mental disorder—though she may still have control over a subset of choices unaffected by the disorder. Perhaps more importantly still, the person remains *capable* of exercising control over her disorder, *if* presented with suitable remediating resources. Her responsibility, in other words, is dormant, not non-existent. This unfortunate scenario is more likely in places like the United States, which are economically polarized and lack universal (mental) health care, compared to more egalitarian cultures.

In cases of "dormant" responsibility, is no one responsible? There is a case to be made that political agents can be held responsible for failing to mitigate social risk factors in their electoral districts, and for failing to provide access to mental-health resources to their constituents. This permits us to blame politicians for high incidents of mental illness in their electoral districts and for *specific incidents* of mental illness caused by socioeconomic deprivation. In other words, when circumstantial factors defeat a person's responsibility, we can hold political agents responsible for causing, facilitating, or enabling conditions

of social injustice, and for the effects of this situation. This "displacement" of responsibility from the primary agent (who directly performs the action) onto the secondary political agent is supported by any theory of responsibility that recognizes responsibility for omissions (e.g., Fischer 2006; Holroyd 2012) or collective responsibility for individual choices (e.g., List and Pettit 2011).

In sum, when assessing a service user's responsibility status, we need to take into account two layers of considerations: (1) the symptomatic profile and response rates of the person's disorder, and (2) the person's psychological profile and her role in a system of relationships, including relations of inequality that may impair agential functioning. The second set of considerations allows us to resist totalizing and stereotypical attitudes toward services users, instead treating them as multidimensional people with unique personalities, circumstances, and communities.

Concluding remarks

Responsibility theorists have a tendency to assume that mental illness and responsibility are mutually exclusive, though this is contradicted by the dominant client-centered approach to therapy and *DSM*-5's definition of mental illness as a multitude of states with variable individuating characteristics and presumed treatment-responsiveness. As an antidote to the blanket exclusion clause, I have proposed an agency-enhancement theory of responsibility that is compatible with client-centered therapy and also respects the robust diversity of service users. This view allows us to hold service users responsible for a substantive range of choices, and to see treatment as a type of reactive attitude rather than a mode of exclusion and isolation. This expands the scope of responsibility to include the vast majority of service users.

Notes

1 I borrow the term "blanket exclusion clause" from Freyenhagan and O'Shea (2013: 57), who use it to describe theories of autonomy that present autonomy and mental illness as mutually exclusive. I apply the term to parallel theories of responsibility: theories that treat mental disorders and responsibility as incompatible.

2 It should be noted, however, that the only time that Sher mentions mental illness in his book, he says that his theory should be taken to exempt "the severely mentally

retarded (*sic*) and the insane" without qualification (2009: 110). Perhaps he is thinking only about people with "shallow" (transitory) disorders. But if he means to exempt all types of mental disorders, he is committed to a Strawson-like blanket exclusion clause. Both approaches are unacceptably generalist.

3 Although Vargas describes AEV as "modular" and compatible with multiple ethical frameworks, Victoria McGeer (2015) rightly points out it is hard to imagine how AEV could possibly be congruous with a deontological (non-consequentialist) framework. I will simply assume that AEV is a consequentialist approach whose aim is to promote agential functioning.

4 Caruso (2016) similarly provides a therapeutic ("public health") approach to responsibility, but Caruso's view differs from mine in the details.

5 I am using Hare's affective-behavioural definition of psychopathy, which is only one amongst many diagnostic classifications. My comments here do not apply to alternative definitions of psychopathy.

References

American Psychiatric Association (2013), *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Arlington, VA: American Psychiatric Publishing.

Arpaly, N. (2015), Huckleberry Finn Revisited: Inverse Akrasia and Moral ignorance. 141–156. In *The Nature of Moral Responsibility: New Essays*. R. Clarke, M. McKenna, and A. Smith, eds. Oxford: Oxford University Press.

Babiak, P. and Hare, R. D. (2006), *Snakes in Suits: When Psychopaths Go to Work*, New York: Regan Books.

Bell, M. (2013a), The Standing to Blame: A Critique, pp. 263–281. In *Blame: Its Nature and Norms*. D. J. Coates and A. N. Tognazzini, eds. New York: Oxford University Press.

Bell, M. (2013b), *Hard Feelings: The Moral Psychology of Contempt*, New York: Oxford University Press.

Blumer, D. (2002), The Illness of Vincent van Gogh, *American Journal of Psychiatry*, 159: 519–526.

Caruso, G. D. (2016), Free Will Skepticism and Criminal Behavior: A Public Health-Quarantine Model, *Southwest Philosophy Review*, 32:25–38.

Clarke, G., Hops, H., Lewinsohn, P. M., Andrews, J., Seeley, J. R., and Williams, J. (1992), Cognitive-Behavioral Group Treatment of Adolescent Depression: Prediction of Outcome, *Behavior Therapy*, 23: 341–354.

Coccaro, E. F. and Kavoussi, R. J. (1997), Fluoxetine and Impulsive Aggressive Behavior in Personality-Disordered Subjects, *Archives of General Psychiatry*, 54:1081–1088.

Cooper, B. (2005), Immigration and Schizophrenia: The Social Causation Hypothesis Revisited, *The British Journal of Psychiatry*, 186(5): 361–363.

Darwall, S. L. (2006), *The Second-Person Standpoint: Morality, Respect, and Accountability*, Cambridge, MA: Harvard University Press.

Davidson, K. M. and Tyrer, P. (1996), Cognitive Therapy for Antisocial and Borderline Personality Disorders: Single Case Study Series, *The British Journal of Clinical Psychology*, 35:413–429.

Dolan, B., Warren, F., and Norton, K. (1997), Change in Borderline Symptoms One Year after Therapeutic Community Treatment for Severe Personality Disorder, *The British Journal of Psychiatry*, 171:274–279.

Frankfurt, H. G. (1969), Alternate Possibilities and Moral Responsibility, *The Journal of Philosophy*, 66(23):829–839.

Fischer, J. M. (2011), *Deep Control: Essays on Free Will and Value*, New York: Oxford University Press.

Fischer, J. (2006). *My Way: Essays on Moral Responsibility*, New York: Oxford University Press.

Freyenhagen, F. and O'Shea, T. (2013), Hidden Substance: Mental Disorder as a Challenge to Normatively Neutral Accounts of Autonomy, *International Journal of Law in Context*, 9:53–70.

Haji, I. (1998), *Moral Appraisability: Puzzles, Proposals, and Perplexities*, New York: Oxford University Press.

Hare, R. (1990), *Hare PCL-R*, 2nd edition. North Tonawanda, NY: MHS Assessments Publishing.

Hare, R. D. (1993), *Without Conscience: The Disturbing World of the Psychopaths among Us*, New York: Pocket Books.

Hare, R. D. (1999), *Without Conscience: The Disturbing World of the Psychopaths among Us*, New York: Guilford Press.

Holroyd, J. (2012), Responsibility for Implicit Bias, *Journal of Social Philosophy*, 43:274–306.

List, C. and Pettit, P. (2011), *Group Agency: The Possibility, Design, and Status of Corporate Agents*, Oxford: Oxford University Press.

Martin, M. W. (2012), Psychotherapy as Cultivating Character, *Philosophy, Psychiatry, Psychology*, 19:37–39.

McGeer, V. (2008), Trust, Hope and Empowerment, *Australasian Journal of Philosophy*, 86:237–254.

McGeer, V. (2015), Building a Better Theory of Responsibility, *Philosophical Studies*, 172(10):2635–2649.

McKenna, M. (2012), *Conversation and Responsibility*, Oxford: Oxford University Press.

Mental Health Commission of Canada (2015), Recovery Guidelines Ottawa, ON: Author. ©2015 Mental Health Commission of Canada. URL= https://www.mentalhealthcommission.ca/sites/default/files/2016-07/MHCC_Recovery_Guidelines_2016_ENG.PDF Accessed January 29, 2017.

Rice, M. E., Harris, G. T., and Cormier, C. A. (1992), An Evaluation of a Maximum Security Therapeutic Community for Psychopaths and Other Mentally Disordered Offenders, *Law and Human Behavior*, 16:399–412.

Rogers, C. R. (1951), *Client-Centered Therapy: Its Current Practice, Implications, and Theory*, Boston, MA: Houghton Mifflin.

Rogers, C. R. (1995), *A Way of Being*, Boston, MA: Houghton Mifflin Harcourt.

Sher, G. (2009), *Who Knew?: Responsibility without Awareness*, Oxford: Oxford University Press.

Smith, A. M. (2005), Responsibility for Attitudes: Activity and Passivity in Mental Life, *Ethics*, 115:236–271.

Smith, M., Calam, R., and Bolton, C. (2009), Psychological Factors Linked to Self-Reported Depression Symptoms in Late Adolescence, *Behavioural and Cognitive Psychotherapy*, 37:73–85.

Strawson, P. F. ([1963] 2003), Freedom and Resentment, pp. 72–93. In *Free Will*, 2nd edition. G. Watson, ed. Oxford: Oxford University Press.

Vargas, M. (2013), *Building Better Beings: A Theory of Moral Responsibility*, Oxford: Oxford University Press.

Walker, C., Thomas, J., and Allen, T. S. (2003), Treating Impulsivity, Irritability, and Aggression of Antisocial Personality Disorder with Quetiapine, *International Journal of Offender Therapy and Comparative Criminology*, 47:556–567.

Wallace, R. J. (1994), *Responsibility and the Moral Sentiments*, Cambridge, MA: Harvard University Press.

Waring, D. R. (2016), *The Healing Virtues: Character Ethics in Psychotherapy*, Oxford: Oxford University Press.

Watson, G. (2013), Responsibility and the Limits of Evil: Variations on a Strawsonian Theme, pp. 84–113. In *The Philosophy of Free Will: Essential Readings from the Contemporary Debate*. P. Russell and O. Deery, eds. Oxford: Oxford University Press.

Wolf, S. (2011), Blame, Italian Style, pp. 332–347. In *Reasons and Recognition: Essay on the Philosophy of T. M. Scanlon*. R. J. Wallace, R. Kumar, and S. Freeman, eds. New York: Oxford University Press.

Zimmerman, M. (1988), *An Essay on Moral Responsibility*, Totowa, NJ: Rowman & Littlefield.

17

Philosophers, Psychopaths, and Neuroethics

Matthew Ruble

Introduction

We must confront the fact that mental illness does not in itself automatically exempt persons from moral responsibility. Antisocial personality disorder, often referred to as "psychopathy" (as I shall in this chapter), is a distinct mental illness that, owing to its being defined by diagnostic criteria that are explicitly related to moral norms, is uniquely conceptually positioned as straddling the divide between medicine and morality.[1] As such, psychopathy is a disorder that is helpful to understand the conceptual debate over mental illness, and whether or not it is best understood as a value-free medical disease.

There is also something about the unique cultural preoccupation we have with psychopaths that warrants philosophical attention. We appear to be infatuated with the portrayal of psychopaths in our storytelling (in books, movies, and television), yet we devote relatively few resources to understanding and treating those alleged to be suffering from this alleged illness (see Blair 2005; Levy 2007a; Sinnott-Armstrong and Wheatley 2014). And even if we are to extend sympathy for psychopaths, we should keep in mind the widespread damage these persons inflict on individuals, families (see Noguera 2000), and society at large (Blair 2005; Levy 2007a). This chapter is not about steering a just amount of health care resources toward psychopaths (after all, perhaps our criminal justice and mental health treatment resources are better served elsewhere). Rather, the general aim of this chapter is to provide a philosophical analysis of the collision of two seemingly irreconcilable ideas: mental illness as understood in a neurobiological era of psychiatry on the one hand, and moral responsibility on the other. What sense can we make, if any sense at all, of the question: "What does it mean to suffer from a *moral illness*?" Is it permissible for us to blame someone for contracting an illness that renders her a morally blind agent? Can

we even reconcile this odd pairing of concepts? Like it or not, psychopaths force us to rethink the boundaries of moral responsibility. Recent developments in the neuroscience of psychopathy provide the advantage of adding empirical evidence to our inquiry (Levy 2007a). But empirical observations alone are insufficient for reaching a verdict on the moral responsibility of psychopaths. We will also need to examine our basic understanding of what it means to be an agent with a capacity for moral responsibility. Fortunately, philosophy has the remaining resources we need to, if not render a verdict on psychopaths, at least greatly increase our understanding of the nature of moral responsibility. And this endeavor has important payoffs for philosophy, psychiatry, and the emerging field of neuroethics.

The conditions for moral responsibility

Aristotle articulated two conditions for moral responsibility that we maintain today. The first is the volitional condition, which inquires into whether or not a person possesses sufficient voluntary control over her actions in order to qualify her as morally responsible for those actions. The second is the knowledge condition, which inquires into whether or not a person possesses sufficient knowledge of her actions to qualify her as morally responsible for those actions. The philosophical literature on the first condition—the voluntariness condition—is *vast*. We will look into one very important essay, contemporary by philosophical standards, written by Peter Strawson, entitled "Freedom and Resentment" ([1962]1993). This essay is foundational for understanding how the voluntariness condition for moral responsibility relates to excuses, and even exemptions from moral responsibility. It is even more important for our purposes because Strawson relies heavily on mental illness as a paradigmatic condition for exempting someone entirely from moral responsibility. And since psychopaths are alleged to suffer from a mental illness that undermines their capacity for moral responsibility, Strawson has provided us with a philosophical gem! But this is not to be taken as an endorsement of his view. For as we shall see, Strawson fails to imagine how we might possibly *morally wrong* a psychopath by exempting him from our moral reactions to his wrongdoing.

Considerably less attention has been paid by philosophers to the epistemic, or knowledge condition for moral responsibility (see Sher 2009). And by extension, the knowledge condition as it pertains to the responsibility of psychopaths, has likewise received little attention. That was true until Neil Levy's (2005, 2007a, b,

2010) robust defense of the psychopath as unqualified for moral responsibility owing to his inability to guide his action with *specifically* moral knowledge. And Levy crucially enlists empirical evidence from neuro imaging (fMRI) to support his argument, thus linking the psychopath's (alleged) abnormal neurobiological function to his lack of moral capacities and his (alleged) lack of moral responsibility. We will look carefully at the views of both philosophers. After doing so we will be much better positioned for understanding how the debate over psychopaths represents a paradigmatic debate over the nature of mental illness itself, and we will be able to begin assessing what, if anything, neuroscience can tell us about morality.

Strawson's legacy and moral framework

This section charts key distinctions drawn by Strawson and leaves us with a conceptual framework for when it is appropriate, or even possible, to hold a moral agent responsible for wrongdoing. With this conceptual framework in place we will be better positioned to see the range of appropriate moral responses available to those wronged by psychopaths. We will also be able to slowly bring into focus just what it is that we are reacting *to*, what our reactive attitudes are alleged to be targeting, and why psychopaths appear to be such confounding and evasive targets for our reactive attitudes. Later in the chapter, both Strawson's framework and Levy's argument will be carefully scrutinized, and alternatives will be offered.

The primary distinction drawn by Strawson is that between the reactive attitudes and the objective view. The reactive attitudes are the attitudes and feelings we have toward others when our interpersonal encounters are morally significant. We feel resentment, or anger, when others wrong us in a morally significant way. We also feel gratitude when others aid us, and when appropriate, we feel sympathy for others when they face unfortunate circumstances. The range of reactive attitudes is both conceptually and emotionally immense.[2] The objective view, by contrast, is an "attitude" far removed from the reactive attitudes. The objective view is partially prompted in us when the reactive attitudes are inappropriate, for instance, when someone has a good excuse that accounts for the moral wrong done (e.g., the offender is merely a small child, or suffers from severe cognitive impairment, or a person fully disoriented in a haze of dementia). The emotionally detached objective view is fully prompted in us when we wonder if the wrongdoer even qualifies as a moral agent at all. If she does not, then she is,

according to Strawson, exempt from all reactive attitudes and hence the objective view becomes for the responder, a "relief from the strains of involvement." Thus, much like a natural scientist in the laboratory or the field, the objective view is emotionally detached, dispassionate, and exceedingly rational.

The distinction between the reactive attitudes and the objective view is interrelated with the distinction between excuses for our wrongdoing, and outright exemptions that remove us entirely from moral responses. Our reactive attitudes are suspended or otherwise diminished in circumstances in which the wrongdoer is excused for her otherwise morally offensive actions or attitudes. We are all familiar with any number of circumstantial excuses for wrongdoing (e.g., the offender was under an extraordinary amount of stress, or coerced, or the offender harmed someone while in the throes of a seizure). When we excuse the wrongdoer we acknowledge that her otherwise normal morally considerate character was temporarily strained owing to circumstances. We know that she would not normally act in the offensive way she did. We know that her moral agency is otherwise intact. Exemptions from wrongdoing are far more sweeping. When we exempt someone entirely from our reactive attitudes we are deducing that the person's moral agency is not intact, or, in the case of the psychopath, moral agency is entirely missing. Since the reactive attitudes are modes of addressing moral agents, when we operate from the objective view we no longer perceive that we are engaged with a moral agent. Strawson's language here is rather striking. When we take up the objective view we are no longer able to morally reason with or hold our reactive attitudes toward an agent. We view that agent as something to be "controlled," "treated," or "manipulated," and for whom the normal terms of moral address are inappropriate. With the objective view comes the suspension of the notion that we are engaged with another moral agent, and if we are engaged in controlling or manipulating this "other," then we are either in moral violation of that moral other, or we are somehow justified (morally and epistemically) in engaging the other as a non-moral other. Later in the chapter we will draw distinctions between the manner in which we experience our attitudes (*reactans*) internally and how those attitudes ought to manifest outwardly.

Strawson refers to moral agents we excuse from moral responsibility in a particular instance as "type 1" agents, whereas we exempt entirely from responsibility "type 2" agents. The paradigmatic "type 2" agent described by Strawson suffers from some manner of mental illness. The "person in a hopeless delusion," and the "psychopath" *qua* "moral solipsist," who "altogether lacks moral sense," and whose "neurotic behavior is determined in a sense that deems the objective attitude is deemed appropriate." Such agents are not those

toward whom the normal terms of moral address and the reactive attitudes are appropriate because we (allegedly) do not see them as "a morally responsible agent, as a term of moral relationships, [or] as a member of the moral community."

Drawing from psychoanalytic models of mental illness Strawson portrays the type 2 agents' neurosis as entirely undermining her moral agency to the extent that she is incapable of controlling her actions to meet the basic demands of morality. Thus, she fails to meet the voluntariness condition for moral responsibility. On this picture, psychopathology wrests from a person her ability to guide her actions based on moral reasons, and with it, revokes her membership in the moral community. And since type 2 agents are not moral agents, the reactive attitudes are no longer appropriate. Although Strawson use of "appropriate" appears elastic, one assumes that he not only means that reactive attitudes toward the severely mentally ill are likely to be an ineffective means of altering behavior but moreover, that when full members of the moral community engage in resenting and blaming the wrongdoing type 2 agent we are engaged in something epistemically *and* morally inappropriate. Our passionate (re)directing of our attitude of resentment toward the neurotic and broken "self" of the type 2 agent is, on Strawson's view, wrong. By default this leaves us with taking up the objective view and its detached attitudes whence we no longer engage the type 2 agent as someone capable of moral address, and rather as something to be "manipulated" and "controlled." And herein lies the irony of the psychopath's inclusion among type 2 agents. Since psychopaths are masterful manipulators (and manipulating others remains a diagnostic criteria) who do not extend even basic moral consideration to those they con and manipulate, then we appear to be left with a prescription from Strawson to simply meet the psychopath on something of his own terms. For some of us this irony is too bitter. The psychopath appears entirely unmotivated to belong to the moral community and our affirming his non-membership (albeit dispassionate and morally cold) ought not be the final say on the matter. Well, for better or worse, philosophers have remained interested in psychopaths if for no other reason than to demarcate the outer boundaries of morally responsible agency.

Gary Watson and the case of Robert Harris

For an example of the way our moral judgments are shaped (and often divide), we now turn to the case of Robert Harris as discussed by philosopher Gary Watson in his influential essay "Responsibility and the Limits of Evil." Watson's

paper is an attempt to address some criticisms of Strawson's essay. It is important to keep in mind that both Watson and Strawson are involved in a much more abstract metaphysical question regarding whether or not *anyone* is free enough to ground moral responsibility. Both think that we are. But both think that some among us—those with severe mental illness, or "misfortune in formative circumstance"—are not free in a morally significant sense. Our proneness to the reactive attitudes and our commitment to keeping those attitudes alive and central to our moral lives appear to be impervious to crafty metaphysical arguments attempting to demonstrate how determined our lives and actions are by something external to our "self." For better or worse our commitment to the reactive attitudes endures. But like Strawson, Watson does think there are persons toward whom we ought to revise our reactive attitudes because they are not free in a morally significant sense, and he makes use of the case of Robert Harris (by now a rather famous case among moral philosophers).

Watson draws from an *LA Times* article about Robert Harris, a real man executed in California (1987) for murdering two teenage boys. Harris's biography unfolds in reverse order in this article and our reactive attitudes follow along. The Harris we first meet is a grown man condemned by the state and reviled even by his fellow death row inmates. Harris found his way to death row by way of auto-theft and the assassination of two teenage boys. His younger brother Daniel served as a direct witness, commenting on Robert's post-murder callous laughter as he finished the partially eaten cheeseburger that remained of one of his fresh victims as he wiped the barrel of his pistol clean. The older Harris even entertained himself with the fantasy that he and his brother dress as policemen to inform the boys' parents of their sons' dual murders, simply for the joy of being entertained by the bereavement they would be sure to witness. The sheer sinister details of Robert's actions that afternoon leave the reader with a reactive attitude of resentment ("antipathy" in Watson's words), *toward* Harris. We see the point of his death row inmates claim that "the dude is a total scum bag." Regardless of our moral stance on capital punishment, most of us feel a negative attitude toward Harris for a constellation of actions that he fully intended and performed with no moral regard for his victims. Furthermore, by way of empathy, most of us also can vicariously imagine the attitudes of shock and grief that the parents must have felt upon hearing the news. Indeed, we think, our reactive attitudes of resentment and disgust are well justified.

We do not know this from the newspaper article, but suppose we come to learn that Harris is a psychopath. This, I think, does not require us to significantly stretch our imagination.[3] If Harris really does suffer from an illness

that undermines his capacity to act as a moral agent, then we are reminded of Strawson's view that requires we diminish our resentment we feel toward Harris. But some of us remain unmoved from our strongly held attitude of resentment simply by learning that Harris has been diagnosed. Are we wrong to remain unmoved? This is not yet clear. And neither have we heard Harris's full story.

After reading about the adult Harris and feeling resentment and antipathy toward him we then learn just how misfortunate he was in his formative circumstances. "Violence presaged his birth," as his father kicked his seven-month pregnant mother such that Robert was thrust into the world, and an idiosyncratically deplorable situation, two months prematurely. Robert suffered bad circumstantial moral luck, if nothing else, as he endured a childhood filled with emotional and physical abuse that ended with an extended sentence in juvenile prison that changed Robert fully into the morally ill man he became. Robert was physically and sexually assaulted and attempted suicide regularly during his juvenile sentence. Following his sentence he emerged a ruined man that his sister describes as "having no feeling, no remorse, and there was nothing much left of him." Aside from the reactive attitudes that are drawn toward Harris's parents, or our lousy retributive criminal justice system, our attitudes toward Harris give way from "antipathy to sympathy," in Watson's words. We feel sympathy for the young Harris and perhaps do not even want to imaginatively feel what he endured in juvenile prison. We are morally sad for Robert the boy. So what are we to do with our competing reactive attitudes toward Robert Harris? And what *should* a diagnosis of a mental disorder provide by the way of guiding our reactive attitudes? Perhaps we need to know more about the idea that mental illnesses act as causal agents that usurp or bypass our true self in such a way that exempts us from moral responsibility. For if a given mental illness indeed undermines my moral self and my ability to guide my actions based on moral reasons, then I indeed would not meet the volitional condition for moral responsibility. Perhaps Strawson has a point.

Before we depart this part of our discussion I wish to note that some of us are unlikely to experience mitigation in our reactive attitude toward Harris the adult when we are only provided with the information that "he is a diagnosed psychopath." Our resentment does not waiver. But for many of us, we must acknowledge, when we hear that Harris's illness means he has a "chemical imbalance," we are more likely to mitigate our resentment by way of our belief that these imbalances undermine our ability to control our behavior.[4] Such a development traces our progressing models of mental illness, and where

Strawson viewed the mentally ill *self* as out of control by way of unconscious drives, our contemporary views see the mentally ill *brain* as out of control owing to neurological dysfunction.

This is true of neuroethicist, Neil Levy, whose compelling defense of psychopaths is as carefully argued, as it is empirically informed. Prior to looking at Levy's view, I want to leave this section noting that many readers did experience a mitigation of resentment toward Harris and not because of a lurking self-undermining diagnosis of a mental illness, but rather his narrative, his life story, elicits our sympathy. Hence, our reactive attitudes can steer toward any number of morally significant features, quite aside from Harris's "self" or "brain." We will take up an expanded notion of moral responsibility later in the chapter but we must note that Strawson (and so too with Levy) think that our attitude (resentment) should not be voiced (blame!) toward Harris (broken-self). "Is" Harris responsible must first be answered (and can only be answered by investigation of his self, agency, and sanity) before we go on to "hold" him responsible (by feeling attitudes, voicing them, and bringing sanctions to bear). But this might be an overly narrow conception of moral responsibility, and one tendency that results from this narrow conception is that we are compelled to gaze deeper into the "is responsible" question, and thus into Harris's self, and where technology permits, gaze into his brain as well. Enter neuroethics.

A new defense of psychopaths: Neuroscience and the knowledge condition

The fact that we members of the moral community experience wide-ranging and diverse reactive attitudes may merely reflect a psychological or sociological observation, which may not in itself be a well-considered or sufficient normative conclusion regarding how we *ought* to respond to wrongdoing committed by folks like Harris. Aristotle long ago noted that we can feel the entirely wrong sentiment in a given situation (anger instead of righteous indignation), and we can feel the appropriate sentiment but in the wrong amount (rage instead of frustration, or clinging to resentment long enough to sour ourselves entirely). Some standards of moral sentiments and moral address do seem necessary for a moral community and we do not want to depart our analysis with only a superficial conclusion that appreciating diverse and plural moral responses is tantamount to an "anything goes" standard. Perhaps empirical evidence can help resolve our divided reactive attitudes toward psychopaths

like Harris. Models of psychopathology have changed since the psychoanalytic times during which Strawson wrote his seminal "Freedom and Resentment." The reigning model has decidedly turned from an understanding of mental illness as one's toxic unconscious drives undermining the moral agency of one's conscious self to that of an exceedingly scientific neuro-biological account that explains our failures of moral agency at such a reduced anatomical level that highly specialized technical equipment is needed to observe our neural dysfunctions.[5] What then does neuro imaging (fMRI) have to offer the debate over the moral responsibility of psychopaths? Enter Neil Levy. When we bring neuroscience to bear on questions of morality, including moral agency, then we are engaged the study of *neuroethics*. Philosopher Adina Roskies refers to neuroethics as the twofold study of the neuroscience of morality and the morality of neuroscience.[6]

Levy (2007a) offers an empirically informed argument that psychopaths lack the capacity for moral responsibility because they fail to meet the knowledge condition for moral responsibility. It is hence wrong to blame them. Let's review Levy's compelling argument that psychopaths are not morally responsible agents in two separate stages, beginning with the knowledge condition argument, and then taking up the addition of neuroscientific evidence of a dysfunctioning ventromedial prefrontal cortex (vmPFC) in the brains of psychopaths that supports his conclusion.

Recall that Strawson's analysis addresses whether or not psychopaths (and other type 2 agents) possess a *self* sufficient to meet the volitional condition for moral responsibility and he returns a negative verdict. But what of the knowledge condition for moral responsibility? Do psychopaths fail to meet that condition as well? According to Levy psychopaths do not meet the knowledge condition and thus, are not morally responsible agents.

I take the following to be a faithful reconstruction of Levy's (2007a) knowledge argument.

1. Moral knowledge is necessary for moral responsibility.
2. The ability to make the moral-conventional distinction is necessary for moral knowledge.
3. Thus, the ability to make the moral-conventional distinction is necessary for moral responsibility.
4. Psychopaths cannot make the moral-conventional distinction.
5. Thus, psychopaths lack moral knowledge.
6. Therefore, psychopaths lack moral responsibility.

The ability to make the moral-conventional distinction in the second premise is doing considerable work in this argument and requires explanation. This is the low-tech portion of Levy's argument that references studies probing perceived sources of moral authority. The principal study cited by Levy uses primary school children as the control group. The children are asked two sets of questions that distinguish merely conventional wrongs (those alleged to revolve around sociocultural norms) from moral wrongs (those alleged to revolve around some source that transcends mere sociocultural norms—and this is unclear). The first set asks the children (an admittedly politically awkward question) if it is ok for boys to wear dresses to school, to which the children respond "no." "What if your teacher said it is not wrong for boys to wear dresses to school? Is it permitted then?" Here the children reply, "yes." So, dress wearing maps onto (so-called) conventional norms, over which the teacher has authority in the eyes of the children. The second set of questions asks the children if it is wrong to pull the hair of a classmate, to which the children respond in the affirmative. "What if your teacher said it is not wrong to pull the hair of your classmate? Is it permitted then?" Here the children reply, "no." Hair pulling maps onto moral norms over which the teacher does not have authority in the eyes of the children. Hence, the children possess the ability to make the moral-conventional distinction. Psychopaths do not, however, and appear to think that if an authority permits hair pulling then there is nothing morally wrong with doing so. Hence, Levy deftly notes, all wrongdoing is merely conventional wrongdoing in the eyes of the psychopath. And the remainder of the argument follows: psychopaths fail to meet the knowledge condition for moral responsibility, and are thus, not morally responsible agents. Such is Levy's conceptual argument from moral knowledge. Let's now turn to the supporting empirical evidence and look at the brains of psychopaths. Rest assured, later we will also look at the way we look at the brains of psychopaths.

Although much attention has been paid to an allegedly undersized amygdala in the brains of psychopaths, there is inconclusive evidence that the size of amygdala correlates to the instrumental violence (as opposed to reactive violence) exhibited by psychopaths (Blair, Mitchell, and Blair 2005; Levy 2007a). Our focus has now turned to the vmPFC as the region of the brain in which moral reasoning is believed to occur.[7] When we observe brain scans of psychopaths as we ask them to consider moral scenarios we do not observe a signal indicating activity in the vmPFC region as we do with non-psychopaths.[8] In fact, Sinnott-Armstrong and Wheatley (2014) notes that there is activity in an entirely distinct region of the brains of psychopaths (right frontal temporal

parietal lobe, RFTP). This region of the brain is associated with our attempts to discern the mental contents of others, in other words, when we wonder what someone else is thinking or desiring. Some call the RFTP the "theory of mind" region of the brain. (We will return below for some concerns relating to this observation.)

What does the neural functioning of the vmPFC have to do with the moral responsibility of psychopaths? The argument appears run the following way. The vmPFC is the region of the brain correlated with moral reasoning. Since the vmPFC of psychopaths does not signal when considering moral dilemmas, the psychopath appears to have a dysfunctioning vmPFC that undermines his ability to morally reason, including his capacity for moral knowledge, hence psychopaths lack the capacity for moral responsibility. It is hence wrong to hold psychopaths responsible by blaming them for moral wrongdoing. Case closed? Not quite.

The implications that neuroscience has for our understanding of morality are not limited to the study of psychopaths. Other neuroscientific research focusing on otherwise healthy brains considering moral dilemmas and reaching moral verdicts shows a range of neural activity that map onto distinct normative moral responses, and these norms divide along deontological and welfare-maximizing lines (Berker 2009; Cushman and Young 2009; Greene 2008). For instance, Cushman and Young (2009) found that when subjects are asked to respond to moral dilemma thought experiments like the infamous Trolley Problem (do we push the lever to redirect the trolley toward one person in order to save five others?) and Footbridge Problem (do we push a person off a bridge to redirect the trolley in order to save five others?), distinct and competing neural mechanisms are paralleled by distinct and competing normative moral theories. A key link between Cushman and Young's observations and that of Neil Levy's (2007a) defense of psychopaths is that of Koenigs, Young, and colleagues' (2007) study of individuals with damage to the vmPFC who reliably reach welfare-maximizing verdicts in moral dilemma thought experiments. Key to Levy's argument that psychopaths are not morally responsible is the role of their non-functioning vmPFC. But psychopaths and those with damage to the vmPFC are not the only ones to reach welfare-maximizing verdicts. Thoroughgoing utilitarians reach such verdicts too but we would be wrong to suggest that the welfare-maximizing utilitarian is *ill*. What is the upshot of this? Even for otherwise healthy brains we appear to have a diversity of vmPFC function that correlates to the ways some normative moral theories divide in the verdicts reached. We might well disagree with the welfare-maximizing utilitarian's decision to push the large person off the

footbridge in order to save five others, but when we do so, we do so for normative moral reasons, and not by appeal to the manner in which our vmPFC appears to be functioning as it does or does not participate in our moral deliberations. And though we might say the welfare-maximizing utilitarian is *morally wrong* to push, we cannot infer that she is *pathological*. And this is precisely what Levy appears to infer about psychopaths when he assumes their vmPFC is dysfunctional: hence, on his view, the psychopath is *pathological*, but not *morally wrong*. Furthermore, the observation that the RFTP region of the psychopath's brain is active when engaged in moral dilemma decision-making can be accounted for. We might wonder when and how psychopaths are gathered as subjects for neuroscientific study: prison. And since the psychopathic test subjects are likely to perceive the test officials as authorities, then we can plausibly account for the "other minds" RFTP region signaling as capturing the attempt of the test subject to sort out what he thinks his authoritative question-asker wants to hear. And this would likewise account for the failure of psychopaths to capture the moral-conventional distinction. *All* transgressions to the psychopath, Levy reminds us, are merely conventional transgressions, the norms of which are handed down by authority that has the power to punish.

To summarize the current dialectical situation, we now have accounts of both the volitional condition and knowledge condition for moral responsibility that appear to demonstrate how it is that psychopaths lack the capacity for morally responsible agency. We have raised some speculation about the credence we should lend to the neuroscientific evidence alleged to support the conclusion that psychopaths, owing to vmPFC dysfunction, are not morally responsible agents. But so far our investigation has focused principally on the question, "*is* the psychopath responsible?" with the assumption that once this is answered we can then move on, if necessary, to how we might "*hold*" psychopaths responsible for wrongdoing. In the following section we will conceptually map out these two questions of "is" and "hold" responsible in order to reveal additional understandings and options for *how* we might hold psychopaths responsible—or at least begin to do so.

On holding responsible

So far by way of informing our responses to psychopathic wrongdoing with any operational detail, what we glean from Strawson is that we should not feel or express resentment toward the psychopath. So too for Levy, but we

also get more operational detail from him. "Holding psychopaths responsible imposes moral costs [on psychopaths]. It requires us to treat [psychopaths] in ways they do not deserve," Levy claims. On his view, when we blame psychopaths we are doing little more than adding to their misery (presuming psychopaths are miserable). But Levy also recognizes that these are dangerous people, therefore "we can excuse a psychopath for wrongdoing without being committed to releasing her to offend again." On this view, we can indefinitely jail psychopaths, but not on grounds of desert or retributive justice, and only on grounds of quarantining or preventing violence. But indefinite imprisonment sure appears to many of us as an attempted form of "holding responsible," regardless of whether our stated motive is to seek retribution from the psychopath for his wrongdoing or we merely act to quarantine his damage, thus taking up the objective view and engaging the psychopath as if he were a mere contagious virus. Both Strawson and Levy appear to understand "holding responsible" merely to mean the expression of blame targeted at the wrongdoer's moral self. And since the psychopath is alleged to have no moral self at all (or a properly functioning vmPFC) then we are justified if not obligated to engage him as an agent of moral address but rather as a non-moral object. As previously noted, this common view represents one rather narrow account among a wider range of options of how we might alternatively conceive "holding responsible."

I wish to note a series of distinctions that, when on display, open a wider range of responses to psychopathic wrongdoing (and *all* wrongdoing for that matter) beyond the rather narrow conception held by Strawson and Levy (and many others). First, we can distinguish *reactandum* (the things being reacted to, or the target of our reactions) with *reactans* (the reactions themselves). Within each of these lie further distinctions.

We can distinguish *holding* the reactive attitudes from *expressing* them in our practices of holding others responsible. When we speak of *holding* a reactive attitude we are describing an internal (intrapsychological) state that is some kind of fluid combination of doxastic attitude (belief) coupled with an emotional attitude. Our belief that we have been wronged corresponds with some emotional attitude, and vice versa. (Unfortunately, resentment is our paradigmatic blame-related attitude.) Though often a bare reflection of the internal attitude we hold, the *expression* of the attitudes *can* be markedly different from the *holding* of an attitude. Perhaps our expressions *should* be different from our holdings. I will first explain the ways these can differ before defending the claim that they should differ, specifically with respect to psychopaths.

When drawing this distinction, in both theory and practice, we must maintain vigilant awareness of the gap that occurs between the reactive attitudes we hold internally and how they are revised when expressed. We must balance the practical gains we seek in our expressions with remaining genuine to the attitudes underlying them, including those attitudes that are affectively rich.[9] We need not be overly concerned, however, that the wider the gap between holding and expressing the reactive attitudes necessarily entails that we are moving out of the domain of moral responsibility. Translated to Strawsonian terms, being compelled to suspend the affectively rich reactive attitudes and thus take up the affect-suspended objective view toward a wrongdoer does not itself remove or exempt that wrongdoer from terms of moral address. Accordingly, the distinction between holding and expressing the reactive attitudes is specifically important for informing our responses to psychopathic wrongdoing, which might well justifiably include the same affective-rich reactive attitudes we reserve for each other.

A further distinction with perhaps even more importance is one that is drawn by mental health professionals between affective and non-affective reactive attitudes.[10] My use of "affective" here follows the clinical meaning of the term as describing an individual's emotional state and the relative passion or depth of that state as it is outwardly expressed in either verbal or nonverbal behavior. For example, "the anhedonic exhibit a flat affect." More to the point, we have a tendency to engage in affective blame, and this comes in relatively charged degrees. I want to pause at this point to note that on my understanding of the common-narrow view (like that of Strawson and Levy), it is this particular subtype of *reactans* that common viewers take to be the exclusive domain of the language of moral responsibility. So, when common-narrow view instructs that we "should not hold psychopaths responsible," what is meant is that we "should not blame them." On my understanding it is this specific domain of *expressing affective reactive attitudes* that we are warned to avoid. I return to this point shortly to crystallize my objection to the common view that this is an overly narrow approach to holding responsible.

We might issue instead, a *non-affective expression* of a given attitude. To imagine this, we must first purge ourselves of the paradigmatic "go to" attitudes of blame and resentment; otherwise the incoherent idea of non-emotional resentment prevents us from considering alternatives. Rather, issuing moral demands can be done absent the heightened level of affective content that Levy and others assume *must* accompany all form of "holding responsible." Our concern here is not so much with the underlying reactive attitude that

is actually held as it is with the manner that that attitude is expressed *qua* a moral demand. So, the candidate *reactans* we have before us is the non-affective *expression* of the reactive attitudes. Now, what of the targets of these reactions, the *reactanda*?

When we form a reactive attitude (momentarily ignoring the distinctions just discussed above), to what is it exactly that we are reacting? For brevity, let us refer to the "thing to which we are reacting" as the *reactandum*. For understandable reasons, the common view assumes the agent is the obvious and proper *reactandum*. This is also generally true of the wider literature on moral responsibility.[11] However, there are additional *reactanda* that can be accounted for, and though we might identify a lengthy list of these (reasons, choices, personal history, etc.), I wish to focus on behavior as the *reactandum* that is particularly relevant to the debate whether or not we should hold psychopaths responsible.

In order to see the value of the shift from steering our reactive attitudes away from an agent and, in place, toward her behavior, we need to recall a core concern held from the common view that we should not hold psychopaths responsible (specifically, that we should refrain from blaming them) because doing so both harms them and places an unreasonable demand on them. The common view objects to blaming psychopaths because they are alleged to be agents incapable of understanding even basic moral obligations (Strawson, Levy) and because blaming them would harm psychopaths in a way they do not deserve (Levy). These objections raised by the common view arise from a prioritizing of the *diseased agent* and her personal history as the *reactanda*. But an agent's mental illness and her history are not the only relevant *reactanda* that can, and do, inform our reactions. We do, and should, include behavior and an agent's future to the list of *reactanda*. Perhaps behavior should be the *reactanda* that warrants prioritizing. Let's look at the case for doing so.

Suppose for a moment that the common-narrow view is correct and that when we express reactive attitudes to psychopaths we harm them. Suppose also that we are wary of alienating psychopathic agents by taking up a purely objective view toward them (and furthermore, suppose that we are wary of the conclusion that we simply lock psychopaths away indefinitely). Here is where the option of prioritizing behavior as the more apt *reactandum* gains its appeal. For, if our reactive attitudes are centered on behavior then we avoid a direct attack on, or criticism of, the character or agent herself. It is one thing to condemn a person and quite another to condemn behavior. Although this may be true generally, it seems to apply even more so in cases of emerging moral agents,

such as children and adolescents, and among these, I think, we should include psychopaths regardless of how protracted and delayed their moral development appears. Character, including self-perception of our own character, is much more difficult to change or modify than behavior and behavioral patterns, especially when we metaphysically overcommit to the idea that behavioral change can only follow character change. For, when we comment on character (as revealed in the reactive attitudes) we are alleging to comment on something constitutive of an agent, something that is more deeply rooted than behavior, more so patterned behavior, which is alleged to reveal one's character. If, however, we diminish the commitment to direct our reactive attitudes at an agent's character and prioritize behavior instead, then we have more reason for optimism that moral improvement can occur by way of behavioral modification. Rather than seeing character, or to be precise—diseased character—as the source of behavior, perhaps we need to acknowledge that character and behavior actually work in a kind of gravitational tandem. This approach also helps avoid the conclusion that psychopaths suffer from an incorrigible disease of moral character, for which there is no treatment.

When we prioritize behavior as the more apt *reactandum*, then the distinction, previously discussed, between *holding* and *expressing* the reactive attitudes previously discussed gains salience. It just may be a brute fact about us that when we hold a reactive attitude internally that we do so with the *reactandum* of the offending agent firmly in mind. It does not follow from this, however, that when we *express* the held attitude that the expression cannot be redirected from the agent to direct reference to the offending behavior rather than the agent herself, her character.

It may be that *we* need to help psychopaths believe that they do not have permanently damaged moral characters and that wrongful behaviors need not, as a matter of Pavlovian reflex, beget a lifetime of wrongful behavior. Robert Harris confessed in his final hours that when he killed the two boys that he knew then he had made his choice to go to hell. Unlike Watson's experience of reactive attitudes toward Harris, it is this confession that is the first, and perhaps the only, time in reading his story that I felt compassion for him. Whether it arose from a religious commitment, or a moral metaphysical commitment, or an underlying belief that his formative years were so bad that he was psychologically ruined, or what have you, what is saddest about Harris's story is his idea that he was eternally condemned for his behavior. I think this suggests that there was some semblance of moral sense somewhere deep within Harris's moral psychology.

Conclusion

What is so troubling about psychopaths is that, given the commitment to the narrow sense of *reactans-reactanda* pairing assumed by the common view, our typical practices work with most people. To be more specific, our customary practice of merging our affective holding of attitudes with an affective expression of those same attitudes (and with matching magnitude) targeted at agents that is so effective in our uncontroversially normal lives, has zero effect on psychopaths. And since our traditional modus operandi of holding responsible utterly fails with psychopaths, the common view infers from this that we have on our hands a thoroughly morally diseased quasi-agent. Even more problematic is the epistemic and moral weight on which we place the role of the vmPFC in distinguishing the morally ill from the morally healthy. When we do so we conflate pathology with the normative moral domain in which we address each other as moral agents, warts and all. Why should we think that our traditional practices of holding responsible (in the narrow *reactans-reactandum* combination just discussed), is a standard that, when *it* fails, reveals that the problem *must be* with the non-responding agent, and never with the standard practice itself? We should not. Perhaps the case of psychopaths forces us to acknowledge that the commitment to affective blame is a poor standard for holding responsible.

Acknowledgments

I wish to thank the following for helpful and insightful comments on this chapter: John Hardwig, Adam Cureton, David Reidy, David Palmer, Lois Presser, students at Appalachian State University, March 1, 2016, the graduate program in philosophy at the University of North Carolina at Charlotte Parr Center for Ethics, April 20, 2016, and the editors of this volume.

Notes

1 Although I use the terms "antisocial personality disorder" and "psychopathy" synonymously, doing so is not immune from the criticism that a crucial distinction between the two is ignored. The alleged distinction is that psychopaths are an extreme version of those diagnosed with ASPD and the Hare Psychopathy Checklist (Hare 1993) alleges to mark this distinction with psychopaths scoring higher on the checklist than those with *merely* diagnosed with ASPD.

2 If we are honest with ourselves, we must acknowledge that we rehearse only a very limited range of these morally significant reactive attitudes and why it is that for so many of us that resentment becomes the paradigmatic and exalted reactive attitude when we are morally wronged. My speculative point here is merely to acknowledge that our examination of morally responsible agency, including that of psychopaths, ought not overlook the crucial role that our own intrapsychological range of reactive moral attitudes (later in the chapter I call these attitudes "*reactans*") play in the wider philosophical issue of the moral agency in general, including psychopaths. The narrower our range of moral attitudes is then the less sophisticated our notions of moral responsibility shall be. We ought not hold the latter hostage to the former.

3 On my view, any clinically informed reader of the *LA Times* article on Harris can follow his diagnostic "progression" from oppositional defiant disorder and conduct disorder in his early years, right on through his graduation to full-blown antisocial personality disorder. Note: earlier in the chapter I comment on what some might consider a troublesome equivocation of those diagnosed with ASPD to psychopaths. My hunch is that Harris would have scored highly on the Hare Checklist for Psychopathy.

4 This is a hunch that has some empirical support (Knobe and Nichols 2008), from work being done in the emerging field of *experimental philosophy*. Inserting "chemical imbalance" in various thought experiments shared with undergraduate students does result in the view that those with such imbalances are not morally responsible for violent actions. This, of course, merely sociologically surveys various views on mental illness and moral responsibility—and from a rather unrepresentative sample of moral agents in the form of (barely responsible traditionally aged) undergraduate students—than it provides philosophical insight about the nature of morality.

5 And as anyone who has experience working in mental health care will note, no patient is ever diagnosed by way of fMRI. All diagnoses are based on thought and behavior and are assessed in a clinical interview. We ought not promote the misleading notion that people are diagnosed with antisocial personality disorder—or any other psychological disorder for that matter—by way of neuro imagery. When we do so, we promote a false model of mental illness that ignores the many values involved in sorting unwanted human conditions from pathological human conditions.

6 Not only do we bring neuroscience to bear on questions of moral responsibility and mental illness, philosophers are also debating whether or not, and how, neuroscience can help us resolve debates over competing normative moral theories such as utilitarianism and deontology (Greene, Berker).

7 Note that the implications of vmPFC function research are significant for discussions well beyond psychopaths, and also engage expanded debate about whether or not neuroscience can help us select one normative moral theory as superior to another (Berker 2009; Cushman and Young 2009; Greene 2008).

8 See Levy (2007a, b), Sinnott-Armstrong and Wheatley (2014), Blair (2005), Cushman and Young (2009).

9 Perhaps the genuineness to which I refer maintains what Strawson calls the "vital thing" that is contained in our reactive attitudes that points us to moral responsibility, and that which he takes to be unaccounted for when taking up the objective view and considering the social utility of the attitudes.

10 This distinction is common in clinician use. See Hannah Pickard's (2011) use of the terms "affective and non-affective blame" in her approach to working with individuals diagnosed with personality disorders.

11 Consider George Sher's (2008) claim, "If anything about the concept (responsibility) is clear, it is that what any given agent is responsible for is a function only of what *he himself* has done or failed to do" [emphasis added], p. 99. Levy (2010) notes two distinct theories of moral responsibility are generally recognized in the wider literature: attributionism and volitionism. The former focuses centrally on the freedom of the agent and whether or not an act is directly attributable to her, whereas the latter focuses on the freedom of an individual's choice in so acting. With respect to volitionism, we might say that the relevant *reactandum* is the freedom status of a choice a given agent makes, but the assessment of the freedom of choice will result in *reactans* that are directed at the agent. Levy (2007b, 2010), though a professed defender of volitionism, is particularly prone to collapse volitionism into attributionism when discussing psychopaths. For instance, he claims that psychopaths are not responsible because, owing to their moral disease, they do not choose to be the person they are, thus, they cannot intend to act morally, and thus, are not responsible for the things they do.

References

Berker, S. (2009), The Normative Insignificance of Neuroscience, *Philosophy and Public Affairs*, 37:293–329.

Blair, R. J. R. (2005), Responding to the Emotions of Others: Dissociating Forms of Empathy through the Study of Typical and Psychiatric Populations, *Consciousness and Cognition*, 14(4):698–718.

Blair, J., Mitchell, D., and Blair, K. (2005), *The Psychopath: Emotion and the Brain*, Malden, MA: Blackwell.

Cushman, F. and Young, L. (2009), The Psychology of Dilemmas and the Philosophy of Morality, *Ethical Theory Moral Practice*, 12:9–24.

Greene, J. (2008), The Secret Joke of Kant's Soul, pp. 35–79. In *The Neuroscience of Morality: Emotion, Brain Disorders and Development*. W. S. Armstrong, ed. Cambridge, MA: MIT Press.

Hare, R.D. (1993), *Without Conscience: The Disturbing World of the Psychopaths Among Us*, New York: Pocket Books.

Knobe, J. and Nichols, S. (2008), An Experimental Philosophy Manifesto, pp. 3–14. In *Experimental Philosophy*. J. Knobe and S. Nichols, eds. New York: Oxford University Press.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F. A., and Hauser, M. D. (2007), Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgments, *Nature*, 446:908–911.

Levy, N. (2005), The Good, the Bad and the Blameworthy, *Journal of Ethics and Social Philosophy*, 1:1–16.

Levy, N. (2007a), The Responsibility of Psychopaths Revisited, *Philosophy, Psychiatry and Psychology*, 14:129–138.

Levy, N. (2007b), Norms, Conventions, and Psychopaths, *Philosophy, Psychiatry and Psychology*, 14:163–170.

Levy, N. (2010), Psychopathy, Responsibility and the Moral/Conventional Distinction. In *Responsibility and Psychopathy: Interfacing Law, Psychiatry and Philosophy*. L. Malatesti and J. McMillan, eds. Oxford: Oxford University Press.

Noguera, A. R. (2000), Mad, Sad or Bad. Moral Luck and Michael Stone, *Nursing Philosophy*, 1:158–168.

Pickard, H. (2011), Responsibility Without Blame: Empathy and the Effective Treatment of Personality Disorder, *Philosophy, Psychiatry, & Psychology*, 18(3):209–224.

Sher, G. (2008), In Praise of Blame, *Philosophical Studies*, 137(1):19–30.

Sher, G. (2009), *Who Knew? Responsibility without Awareness*, Oxford: Oxford University Press.

Sinnott-Armstrong, W. and Wheatley, T. (2014), Are Moral Judgments Unified? *Philosophical Psychology*, 27(4):451–474. doi:10.1080/09515089.2012.736075.

Strawson, P. F. ([1962] 1993), Freedom and Resentment. In *Proceedings of the British Academy, Perspectives on Moral Responsibility*. J. M. Fischer and M. Ravizza, eds. Ithaca, NY: Cornell University Press.

Watson, G. (1987), Responsibility and the Limits of Evil: Variations on a Strawsonian Theme, pp. 256–286. In *Responsibility, Character, and the Emotions*. F. D. Schoeman, ed. Cambridge: Cambridge University Press.

Study Questions

1. *When do people with mental disorders lack autonomy? Under what circumstances should we say they are not responsible for their actions?*
2. *Should people with mental disorders be coerced to receive treatment?*
3. *How do we give uptake to the experiences of people with mental disorders?*
4. *What are the strategies to overcome stigma in mental disorders?*

Part Six

Social and Political Philosophy and Psychiatry

18

Neurodiversity Theory and Its Discontents: Autism, Schizophrenia, and the Social Model of Disability

Robert Chapman

Introduction

The neurodiversity movement is a grassroots political movement that initially arose in the late 1990s. Its proponents challenge what they take to be the default pathologization and undue medicalization of "natural human variants" by institutional psychiatry as well as society more broadly. In contrast to the earlier anti-psychiatry movement, which often challenged the biological or pragmatic validity of psychiatric classifications, neurodiversity advocates accept that classifications such as autism and schizophrenia can be useful and meaningful ways to group underlying neurocognitive differences—albeit if they are reclaimed from psychiatry by those so-labeled. What they reject is the idea that (at least some of) these kinds are inherently disordered. Rather, for neurodiversity advocates, some psychiatrics kinds, most notably, autism and ADHD, as well as some other conditions such as Down syndrome and dyslexia, are minority modes of neurocognitive functioning that are disabled by a hegemonic "neuro-typical" (i.e., "normal") society.

Since its inception, the movement has grown rapidly, largely online via blogs and forums, and through self-advocacy organizations, such as the Autism Self-advocacy Network. It increasingly influences policy, public perception, and clinical practice. In the United States for example, Ari Ne'emen, founder of the Autism Self-advocacy Network, was appointed by Barack Obama to the National Council on Disability in 2009; while in the UK, the Labour party announced in 2016 that they were working on introducing the role of Neurodiversity Minister as an official governmental appointment. Despite this, in the academic literature there has only been limited interest in what I call "neurodiversity theory," by

which I mean the central theoretical claims and philosophical arguments that the movement is bound up with. That is, there are now many academic works on how neurodiversity theory relates to, say, public policy, education, and literature; but there have been few rigorous and systematic analyses of the concept itself, or of its key philosophical nuances. Indeed, and despite the key claims of neurodiversity advocates being essentially philosophical in nature, philosophers of psychiatry have, oddly, had almost nothing to say on the matter.[1] Perhaps because most of the debate is carried out in, say, the blogosphere rather than the academy, it has not so far been taken seriously within the philosophy of psychiatry.

This chapter seeks to change this. Bearing the lack of scholarly engagement between philosophers of psychiatry and the neurodiversity movement in mind, the aims of the chapter are twofold. First, to introduce the reader to key concepts and arguments of neurodiversity theory as it currently stands. Of course, since the neurodiversity movement has no leader and no textbook, the arguments and claims its proponents forward are heterogeneous. Nonetheless, the most nuanced account (in both the academic and lay person literature) comes from autistic activist and interdisciplinary academic Nick Walker, published on his blog *neurocosmopolitanism.com*. Given this, I shall focus most centrally on his account, and supplement it with others, as a contrast to standard psychiatric theory. Second, I will critically examine the limits of the neurodiversity claim in relation to what psychiatrists consider to be more severe conditions, in particular, so-called severe autism and schizophrenia. Often, in the public debates, such cases are presented as limit cases, which the neurodiversity movement cannot account for. Sometimes such examples are used to dismiss the claims of neurodiversity theory altogether. Against this, I shall suggest that this worry is not obviously founded. That is, although I cannot completely dismiss these worries in this short chapter, I shall at least question the default assumption that these cases are just obviously problematic for neurodiversity proponents. Rather, I will argue, taking neurodiversity seriously, and resolving the dispute, may require delving to a deeper level.

The neurodiversity paradigm

According to Walker, what the neurodiversity movement calls for at the theoretical level is, most fundamentally, a paradigm shift. Following Kuhn (1962), Walker defines a paradigm as "a set of fundamental assumptions or principles, a mindset or frame of reference that shapes how one thinks about and talks about a given subject" (Walker 2014). This is important, for Walker, because, in regard to both

scientific and ordinary observations, a paradigm "shapes the ways in which one interprets information, and determines what sort of questions one asks and how one asks them." So, for example, whichever paradigm they are viewed through will drive which way research into classifications like autism or dyslexia goes, and in turn the results of the research will, due to being interpreted via whichever given paradigm driving it, at least often seem to support or confirm the assumptions of that paradigm. For example, as one recent review of the scientific literature found, due to being viewed through the currently dominant psychiatric paradigm, whether parts of autistic brains are found to be larger or smaller than neurotypical controls, they are, typically, automatically interpreted as relatively deficient (Mottron 2011).

In particular, Walker distinguished between the *neurodiversity paradigm*, which refers to the underlying axioms and principles exhibited by the neurodiversity movement, and the *pathology paradigm*, which he takes to sit behind psychiatric and biomedical theory and practice as it currently stands. He takes this latter paradigm to be based on two fundamental axioms (2013):

1. "There is one 'right', 'normal', or 'healthy' way for human brains and human minds to be configured and to function (or one relatively narrow 'normal' range into which the configuration and functioning of human brains and minds ought to fall)."
2. "If your neurological configuration and functioning (and, as a result, your ways of thinking and behaving) diverge substantially from the dominant standard of 'normal', then there is Something Wrong With You."

In other words, it holds that individuals and groups who fall outside dominant conceptions of normality, to what is deemed a clinically significant level at any given time or place, must be internally pathological. In turn, this leads to seeing such problems as internal medical issues, and into questions regarding how to "treat" and "cure" such deviations.

More specifically, when it comes to psychiatric theory, there is a general consensus that mental disorder exists and that it can be defined in terms of "harmful dysfunction" (APA). To count as a harmful dysfunction in this sense, two conditions must be fulfilled. First, there must be a deviance from "normal" functioning at the neurocognitive or emotional levels of functioning that in turn typically manifests in tendencies toward certain patterns of behavior at the behavioral level. And second, this dysfunction (and its correlating behaviors) must be associated with harm or suffering, either for the person with the condition, or toward others, to what is deemed a "clinically

significant" level. Notably too, these are both judged precisely in relation to what is taken to be a species-standard norm: that is, what is considered normal for the particular species, in this case humans. On the medical model of mental disorder, then, any ensuing disablement or distress is taken to arise, largely at least, from some biological, neurocognitive, or emotional pathology *within* the individual, due to an internal abnormality (i.e., dysfunction). And this is the case even if some social factors were causally relevant, as in the case of some personality disorders. In other words, even though there is still some room here for suffering caused by social stigma and environmental factors, this is taken to be secondary to the primary harmful dysfunction within the individual.[2]

By contrast, Walker takes the neurodiversity paradigm to be based on three different fundamental axioms (Walker 2013):

1) "Neurodiversity is a natural and valuable form of humanity."
2) "The idea that there is one 'normal' or 'healthy' type of brain or mind, or one, 'right' style of neurocognitive functioning, is a culturally constructed fiction, no more valid (and no more conductive to a healthy society or to the overall well-being of humanity) than the idea that there is one 'normal' or 'right' ethnicity, gender, or culture."
3) "The social dynamics that manifest in regard to neurodiversity are similar to the social dynamics that manifest in regard to other forms of human diversity (e.g., diversity of ethnicity, gender, culture). These dynamics include the dynamics of social power inequalities, and also the dynamics by which diversity, when embraced, acts as a source of creative potential."

Regarding (1), just as biodiversity is crucial for the ecosystem to survive and flourish, so too, according to neurodiversity proponents, is neurological diversity similarly crucial for humanity. Autistic thinking, for example, has proven to be very adapted to working with computers and technology (Badcock 2009), even though this seemed unimaginable to most researchers several decades ago. Given this, regarding (2), in contrast to the pathology paradigm, the neurodiversity paradigm takes a much wider range of ways of functioning to be legitimate, normal, and healthy. It is just that some will be more or less adaptive in different contexts, meaning that the context can make minority modes of being *seem* like they are inherently disabled (I'll return to this below).

Regarding the social dynamics noted in (3), Walker proposes a further distinction between *neurodivergence* (or *neurodivergents*), which refers to those persons who fall outside dominant social norms regarding neurocognitive

functioning; and *neurominorities*, which refers to constructed categorizations of neurodivergents (e.g., autistics, dyslexics, and so on). Significantly, when it comes to neurodivergencies, Walker is referring specifically to *pervasive* neurocognitive differences that he takes to be intimately related to the formation and constitution of the self. So, in contrast to something like epilepsy or passing anxiety, categories such as autism and ADHD indicate different ways of being human. Being autistic, for example, as autistic self-advocate Jim Sinclair puts it "colors every experience, every sensation, perception, thought, emotion, and encounter, every aspect of existence" (1993), meaning that autism is not best thought of as a condition that alters a preexisting person, but rather as a different "way of being." Hence, in contrast to speaking of "people with autism" or "people with schizophrenia," neurodiversity paradigm proponents talk of "autistic people" or "schizophrenic people." When it comes to neurominorities, by contrast, even though it is natural for each neurodivergent individual to be as they are, the categories we group them into are taken by Walker, to varying extents, to be social or pragmatic constructions of family resemblance style clusters of neurocognitive traits (of individuals who related oppression and exclusion). This framework allows for categories such as autism, schizophrenia, and so on, for which no biological essence has been found (despite decades of searching), to be understood as minorities in the way that ethnic minorities are considered meaningful (i.e., despite not being real in a biological sense). That is, they share a similar way of being, which is meaningful due to complex historical circumstances. It also opens up the room for talking about, say, autism, as a political identity with its own culture and shared vocabularies, just as we can talk about gay culture despite there being no "gay gene" (Chapman 2016).

Lastly, in contrast to the medical model, the neurodiversity paradigm is intertwined with the social model of disability, whereby the disablement (or harm, distress, or suffering) of minority ways of functioning stems from centrally from social norms and structures. Social model theorists make a distinction between *individual limitations* (or impairments) and *societal disablement*. The central point is that all humans have various limitations (whether physical or mental), but, for the most part, they only become disabled when the norms and structures of society stop them from functioning. Although there are many proposed versions of the social model, and while Walker himself does not commit to any particular version, perhaps the best fit is the "natural human variation model" of disability proposed by Scotch and Schriner. In their words:

> The natural human variant model of disability defines disability as "the extension of the variability in physical and mental attributes beyond the present—but not the potential—ability of social institutions to respond […] In this conception […] the problems faced by people with disability might be seen as the consequence of the failure of social institutions (and their physical and cultural manifestations) that can be attributed to the institutions' having been constructed to deal with a narrower range of variation than is in fact present in any given population." (1997: 155)

So, for example, although a Dyslexic person may, just as everyone else, have some type-specific limitations when it comes to their cognitive processing, they will only be disabled in a society that is structured in such a way that systematically fails to accommodate for their processing. This is taken to account for any disability or distress that dyslexic persons might experience as stemming from social causes rather than being traceable back to individual medical pathology.

Disorder and diversity

How are we to assess which paradigm, if either, is preferable? Given that the difference is, in large part, one of differing fundamental axioms and assumptions, an initial worry might be that rational enquiry has limits when assessing which paradigm is preferable. The worry here regards how different paradigms are sometimes thought to be *incommensurable*: that is, due to sharing different fundamental axioms, being concerned with different questions, and so on, then there is an extent to which they lack a common measure for cross-communication or from which to weigh their merits against each other (Kuhn 1962). While I do think there is something to this worry—as shall become clear later—it is also the case that the two paradigms do share at least some common ground, and so some level of mutual agreement is possible. It is to this common ground that I shall now turn.

I take there to be at least four significant overlaps in this regard. First, both accept that there is at least the possibility of at least one inherently undesirable neurocognitive condition that we simply should attempt to treat or cure. For example, Merriam, whose general line of thought is more in line with neurodiversity paradigm proponents, notes how:

> There are certain circumstances […] that are so extreme as to make flourishing all but impossible. Anencephaly, a developmental disability in which a child is born without a cerebrum, is one such condition. Anencephalic children are

> born without the neural structures necessary to see, hear, feel pain or have any conscious awareness of themselves or their environment at all. There is no cure or standard treatment for children with anencephaly, and those that are not stillborn usually die from within a few hours to a few days after birth. It is, with little doubt, one of the worst imaginable conditions a newborn can have. (2010: 139)

Assuming this characterization is even remotely accurate, this is clearly a neurocognitive condition (even though it is not classified as a mental disorder) in some sense or another. Moreover, it at least seems to fulfill Walker's condition that it must affect the development of the self in a pervasive manner, since this clearly delineates the (lack of) selfhood of the infant.[3] Nonetheless, despite fulfilling the conditions Walker proposes for counting as neurodivergence, this is also, I think undeniably, a case of harmful dysfunction, in some sense of another, which demands a search for a cure. And it seems to me that any neurodiversity paradigm proponent would, although they might prefer to adopt a different vocabulary, accept this. (And indeed, if they would not, then this would be a good reason for not taking their specific understanding of the neurodiversity paradigm seriously.)

The second is a shared concern for minimizing harm, in the sense of suffering or disablement, of those concerned, and, conversely, of cultivating the possibility of their flourishing, in the sense of being able to live good, meaningful lives. That is, both camps are, although they disagree on a lot, concerned with minimizing the distress and disablement faced by those currently given psychiatric labels; and thus there is an important sense in which what both really want is to foster conditions that allow flourishing. It is just that they disagree in regard to how to frame and achieve this.

The third overlap regards how each paradigm allows for at least some level of pragmatic compromise in relation to the commitment to minimizing harm (and cultivating flourishing). For, even if we cannot wholly vary that one paradigm is right and the other is wrong (i.e., in some kind of objective or even scientific sense), the commitment toward both avoiding undue suffering and cultivating flourishing should allow at least some weight to be given to whichever paradigm ultimately works better in relation to these ends, as the preferable option. For example, if one allowed a self-narration that aided the flourishing of, say, autistic people, then this would be counted as a good reason for adopting that paradigm.

Lastly, the fourth overlap regards how pathology paradigm proponents are, like neurodiversity paradigm proponents, also generally committed to not unduly pathologizing difference. From a clinical perspective, undue pathologization occurs when the risk of negative effects seem to outweigh the potential positives

Thus, for example, many clinicians will avoid giving a diagnosis of borderline personality disorder due to a belief that the harm caused by the stigma associated with this label outweighs the potential benefits of giving it, even if they think the label is accurate. Similarly, if they do not think the label "mild ASD" will benefit an individual in terms of their self-narration, and may actually harm them, then they might see this as a good reason to deny an otherwise seemingly accurate diagnosis. So, in line with neurodiversity paradigm proponents, pathology paradigm proponents often are, despite initial appearances, opposed to pathologizing in cases when they take it to be unhelpful.

So both frameworks allow for the notion that there can be some inherently harmful dysfunctions that demand a cure, and also that unduly pathologizing is problematic. And they both take these things to be important due to a general commitment to aiding the flourishing of those we currently label as disordered. It is just that they disagree in regard to both how we should measure such cases, and where the boundaries lie, when it comes to the majority of neurocognitive differences. Really, then, the key questions when it comes to the two paradigms regards who we count as falling outside this standard, where we draw the limits, and how we might measure this. And as we saw above, this difference can be accounted for largely, although not wholly, in terms of how well the models of disablement employed by each paradigm can account for the distress and disablement associated with each, coupled with a further pragmatic commitment to cultivating flourishing and minimizing distress. That is, when looked at in relation to the social and medical models, it may be that we can compare the validity and usefulness of each paradigm in relation to each other in light of which conditions, or neurominorities, each seems best suited to. So clearly, as I have suggested, anencephaly is best accounted for on a medical model, in relation to something like the pathology paradigm, since it extinguishes the possibility of flourishing under any possible society (so far as I can see). But the disablement of other conditions might seem best accounted for by the social model, even if this seems counterintuitive on the face of it.

Most notably, in this regard, what used to be called "Asperger's syndrome" (henceforth "autism") has been established in the literature as an example of neurodivergent functioning that does not obviously bring any inherent disposition toward clinical levels of suffering or disablement. For example, on the one hand, economic ideology, gender norms, and more general neuronormative social norms and structures have been implemented in the disablement and pathologization of those given this label (see, for example, Timimi et al. 2010). And on the other, the cognitive style and capabilities

associated with the label come with strengths as well as limitations; and they also seemingly come with a different rather than deficient kind of moral agency (Baron-Cohen 2011). So the condition really does seem like a difference, rather than a disorder. Nonetheless, in contrast to this well-established case, critics of the neurodiversity movement often present more so-called severe conditions, such as what is sometimes called "low-functioning" or "severe" autism, and schizophrenia. For these critics, such cases are as obviously pathological as I take anencephaly to be. Given this, such critics take them to show the neurodiversity paradigm to be flawed, or at the least very minimal in its application. It is to these two cases I shall now turn.

Multiply disabled autism as neurominority

The condition categorized in the *DSM*-5 (APA) as "autism spectrum disorder" to a "severe" degree is diagnosed in light of what are taken to be profound social and communication problems, accompanied by restricted and repetitive interests, actions, and routines. Notably, even within this specific subdivision (sometimes called "low-functioning") of the autism spectrum, there is a wide range of kinds and levels of functioning. For example, some supposedly "severely" autistic people, such as neurodiversity advocate Elizabeth Grace, can write eloquently and express nuanced levels of understanding of the world and their place in it. Others have been unable to successfully communicate at all, and their condition can also be accompanied by learning disability, epilepsy, seizures, and a tendency toward self-injurious behaviors. I am concerned here with this more complex kind of case: that is, cases where a nonverbal autistic person is also learning disabled, has epilepsy, and needs twenty-four-hour care.

Given these various factors, this kind of multiply disabled autism (as I shall call it, for lack of a better term) is precisely the kind of case which is usually presented as a limit to the neurodiversity paradigm, since it seems undeniable to many people that the capacity for living a good life is, indeed, inherently and severely diminished. In a blog post prominent neurologist Manuel Casanova writes that:

> The neurodiversity movement calls into question whether higher functioning autistic individuals are representative of the autistic population at large. Would proponents of neurodiversity feel the same way if riddled with seizures, self-injurous behaviour, or tremendously diminished cognitive processing? Would they try to find a cure if suffering with any of the aforementioned handicaps? (2013)

This sentiment is echoed in the little academic literature there is on the matter. For example, Jaarsma and Welin (2012) suggest that what they call "the neurodiversity claim" (i.e., that a condition is a difference rather than a disorder) is only justifiably applicable to so-called high-functioning but not "low-functioning" autism. It is also echoed among many parents of such autistic children, who often feel like the neurodiversity movement is dominated by autistic people, rather than people like their offspring.

Significantly, leading autism researcher Uta Frith probably speaks for many when she writes that to deny that there is a disorder in such cases is not just wrong but "perverse" (2008). What I think critics of neurodiversity see as "perverse" regards what is seen as an overlooking or dismissal of the terrible suffering experienced by many multiply disabled autistic persons (and their families). That is, the notion that autism is a difference rather than disorder might be taken to overlook the heightened disposition toward a lower quality of life that appears to come with the condition, and the various kinds of distress that seem to result from this. The worry is, then, partly an ethical one, regarding a purported misrecognition of autistic suffering. On Casanova's understanding, for example, neurodiversity proponents irresponsibly ignore "the existence of seizures, mood disorders, impaired attention, learning difficulties, or sensory abnormalities in the majority of autistic individuals" (2015).

Nonetheless, it is not clear to me that the existence of seizures, self-harm, and so on provides grounds to dismiss the applicability of the neurodiversity paradigm. There are two key replies here. The first thing to note is that the neurodiversity paradigm does not, after all, dismiss or overlook the suffering autistic people typically face. At least on Walker's account, the paradigm explicitly allows for, say epilepsy, or seizures, to be cured. Since these kinds of things are not, after all, part of the autism (i.e., they are different conditions that exist in the same person), or indeed, part of the constitution of selfhood. Given this, these regularly co-occurring conditions can and should, for Walker, be thought of as falling precisely within the domain of medical science. Notably too: neither is the fact that they do regularly co-occur with being autistic enough to conclude that the autism itself is pathological; by analogy, being a woman is causally and statistically associated with a heightened risk of developing breast cancer, and yet we do not thereby conclude that womanhood itself is a disease.

Beyond this, it is worth recalling here that the neurodiversity paradigm fully acknowledges that all humans have inherent limitations. Given this, significantly, there is no reason to think that endorsing the neurodiversity paradigm means

denying that neurocognitive disability requires aid, in varying forms. Thus, it is perfectly consistent with the paradigm to think that such multiply disabled autistic people may require, for example, type-specific forms of education (just as neurotypicals also benefit from type-specific forms of education). This is also the case with sensory abnormalities: neurodiversity proponents do not deny their existence, they just argue that the world needs to be changed to accommodate for them, rather than the other way around (and they also stress how they can sometimes bring positive experiences). Indeed, there *may* even be room for cognitive enhancement, it is just that this would not be driven in light of a drive toward normalization; rather, the ethical issues involved in any given case would echo those that arise when a neurologically typical person considers cognitive enhancement (e.g., whether it is for good reasons given the very specific context, and so on).

Lastly, as to self-injurious behaviors, on the one hand, it is far from clear that they are always caused by autism in any meaningful sense either. Multiply disabled autistic people who are able to communicate (e.g., by typing), and who partake in self-injurious behaviors, have reported that they felt the need to do so due to social reasons, for example, being bullied or not listened to. The idea that self-injury is wholly traceable back to the person doing it as opposed to being an understandable reaction to external events, when looked at from this perspective, seems more like victim-blaming, which is never helpful (Bridget 2012). And on the other hand, even if there are some self-injurious behaviors that in some cases are more intimately intertwined with the selfhood of multiply disabled autistic individuals, it is far from clear that this necessarily makes these individuals uniquely disordered in a clinically significant sense. Consider some more neurotypical self-injurious behaviors. I doubt many, or perhaps any, multiply disabled autistic persons have a smoking addiction, and yet many neurotypical persons regularly partake in this highly self-injurious activity. But of course, despite this self-injurious behavior being something more generally intertwined with neurotypical being than autistic being, we do not thereby think of neurotypicality itself as a disease. So it is not clear that this leap of reasoning should necessarily be accepted when it comes to autism.

The second point worth noting regards the conceptual feasibility of "curing" a person's autism. Bearing the above points in mind, perhaps the key pragmatic difference in terms of what the neurodiversity will not be able to offer to help the suffering of the autistic individual regards attempts to "treat" or "cure" the autism itself. While some may still reject the applicability of the neurodiversity paradigm precisely in the hope that they might one day be able to cure autistic

persons, the problem with this, for neurodiversity proponents, is that it is not clear that this makes much conceptual sense. As Jim Sinclair (1993) notes:

> Autism isn't something a person has, or a "shell" that a person is trapped inside. There's no normal child hidden behind the autism. Autism is a way of being. It is pervasive; it colours every experience, every sensation, perception, thought, emotion, and encounter, every aspect of existence. It is not possible to separate the autism from the person—and if it were possible, the person you'd have left would not be the same person you started with.

The point here, in other words, is that, given that autism is pervasive, in the sense that it hugely affects cognition, perception, and how the self and all its relations develop throughout the whole lifespan (something accepted by proponents of both paradigms), then if an autistic person was "cured," it is far from clear that the person left over would be the same person that existed prior to the procedure. Rather, what you would have seems more like the destruction of one person, and the replacement of that person by another (albeit in the same body). In other words, "curing" someone's autism really means ending the life of the autistic person, in a certain sense, and replacing them with someone else.[4]

At this point, it may still be replied by the pathology paradigm proponent that, given how much lower the various cognitive capacities, as well as autonomy and agency, of multiply disabled autistic persons seem to be, then the search for a cure is still justified. After all, these capacities are standardly seen as something inherent to flourishing personhood, and to living a meaningful human life. So the pathology paradigm proponent might think that the destruction of one person for another is, although horrible, still justified, even morally obligatory. But here we reach a deadlock. Since, for this reply to hold, we have to presuppose precisely that human value should be judged against some kind of species-standard norm, rather than accept that there are different ways of being human, each with their own norms, and potential values and ideals. Indeed, for the neurodiversity paradigm proponent, thinking there is something wrong with the autistic person just because they (at least seem to) have lower cognitive capacities than the neurotypical, would be like thinking there was something wrong with a Labrador in light of it having lower cognitive capacities when compared with a border collie. In other words, it would rely on some kind of pre-scientific conflation between different categories, which in turn leads to the default pathologization of difference. But for the pathology paradigm proponent, it would be wholly right to do this, since they rely on a different pre-scientific presupposed species-standard norm in order to judge the healthiness of any given kind of neurocognitive functioning. That is, they presuppose humanity

to have one center or normality rather than many, and thus take those who both fall outside this and happen to be suffering for whatever reason to be ill. The disagreement may come back, in part, then, to a difference in fundamental axioms and commitments, as well as value-judgments and ideological baggage, presupposed by proponents of each respective paradigm.

Schizophrenia as neurominority

Putting autism aside for the time being, I now want to turn to schizophrenia. Like autism, schizophrenia has no single biomarker, and is diagnosed by self-reports and behaviors. The four key "positive symptoms" it is associated with in the *DSM*-5 are hallucinations, delusions, disorganized thinking (speech), and grossly disorganized or abnormal motor behavior (including catatonia) (APA 2013). These typically manifest during late adolescence or early adulthood. Beyond this, the condition is also associated with negative symptoms, meaning those which occur in early life and prior to the onset of the positive symptoms of schizophrenia. These include blunt affect, lowered social drive, and lethargy.

Typically, the onset of schizophrenia is taken to be a tragic biomedical pathology that destroys or at least radically diminishes the previously relatively normal person. In line with this, Casanova (2013) writes the following:

> Can anyone imagine taking seriously the argument that we shouldn't be trying to cure people who are experiencing hallucinations, but instead, be more accommodating for their "differences," and that instead of looking for [biological] causes, we should be focusing on a suite of nebulously defined "services" for schizophrenics? (2013)

I take Casanova's position to be that seeing schizophrenia as a neurominority (i.e., non-pathological) is just plain absurd. By contrast, Walker and others routinely include schizophrenia in their lists of neurominorities. Interestingly, I have been unable to find any systematic arguments proposed in favor of supporting either Casanova's exclusion of schizophrenia from, or Walker's inclusion of schizophrenia into, the realms of neurodiversity. But this does not mean we cannot explore the possibility here.[5] In short, although Casanova does not given an argument for his dismissal of schizophrenia, I take there to be two key worries in regard to schizophrenia not fitting the neurodiversity paradigm.

The first is the general worry that it may indeed seem that schizophrenia is inherently harmful (i.e., rather than the harm being caused by society); and the second regards its late onset, and how it can seem to change people who

previously appeared relatively "normal" beyond recognition. This second worry regards how and to what extent the condition relates to the personhood of the schizophrenic person: in other words, whether and to what extent we see it as an illness that destroys the previously healthy person on the one hand, or a natural development their way of being, on the other. This might seem more problematic than in the case of autism since, in contrast to it delineating the development of the self throughout the whole lifespan, it may seem, as one man writes of the emergence of his brother's schizophrenia, that "the brother I grew up with had died" (quoted in Glover 2014: 374). The neurodiversity proponent would of course contest this, and say that this was a natural way for schizophrenic persons to develop and perhaps respond to their environment.

Notably too, this is not a purely conceptual issue: its significance stems largely from the extent to which seeing schizophrenia as part of identity is helpful or not for those so-labeled. In support of the notion that schizophrenia can be a genuine part of the person in a helpful way, Jonathan Glover notes how Simon Champ, who is diagnosed with the condition, first experienced his schizophrenia as an external illness, but over time he came to see it as an integral and positive part of his identity. For Champ, coming to terms with being schizophrenic was a matter of a self-communication:

> that has given me the most precious thread, a thread that has linked my evolving sense of self, a thread of self- reclamation, a thread of movement toward a whole and integrated sense of self, away from the early fragmentation and confusion I felt as I first experienced schizophrenia. (quoted in Glover 2014: 386)

As Champ indicates, this eventual realignment with his condition was not just a technicality, but rather something that helped him live a good, meaningful life. This may give some initial support to the neurodiversity paradigm's notion that the condition should, indeed, be seen as an inherent part of selfhood.

But many others do not reach the same self-relation realized by Champ. When it comes to distinguishing between the person and the illness, Glover himself argues that the question of whether the schizophrenia is part of the person or not comes down to two things: "The person's own reflective endorsement, and his or her active autobiographical story" (2014: 377). Endorsement in this sense regards a relatively stable affirmation of the condition that reflects the person's values, while the self-narration is relevant since the current self can be authenticated partly in light of how an "active, self- creative autobiographical story" allows one to make sense of it (2014: 377). If both of these are possible, then there is, according to Glover, good reason to think that any change in a person

is a genuine part of who they are. For Glover, Simon Champ seems be a clear case where someone has integrated their schizophrenia into their personhood, and so has adjusted to it in a healthy and authentic way. Nonetheless, Glover himself still counts the schizophrenia itself as an illness, and stresses that what he calls "recovery" will in many cases also require medication. So he stops short of endorsing a neurodiversity paradigm perspective when it comes to the condition as such.

Clearly, many schizophrenic persons do experience their way of being as highly distressful, and many find partial respite in medication. But does this mean we should necessarily prefer to think of schizophrenia as an illness? Whether this is justified or not may come back to the question of harm: for if the condition is not inherently harmful, then it seems justified to count it as neurodiversity, but if it is, then it seems delusional to deny that it is a disorder. Notably, some have argued that the distress experienced by schizophrenic persons may be accounted for by factors beyond the individual. Louis Sass, for example, argues that both schizophrenia and postmodern selfhood are defined largely by modes of "hyper-reflexivity," alienation, and an ever increasingly decentered and fragmented world. Given this, Sass argues, there is a sense in which postmodern selfhood and society may contribute or amplify the various problems associated with schizophrenic being (Sass 2004: 320). Relatedly, a recent anthropological study on hearing voices in the United States, India, and Ghana also found important differences in regard to the distress associated with the voices. While voices tended to be experienced as invasive and upsetting in the United States, people in Ghana and India were more likely to find them benign, playful, and even helpful (Luhrmann et al. 2014). Notably, the authors hypothesize that the difference in Western and non-Western experiences of hallucination stems from alternative conceptions of selfhood. In the United States, an individualistic, atomized conception of self dominates, leading to voices seeming invasive and as a sign of lacking self-control. By contrast, in Ghana and India, a more collectivist conception of self dominates, and so "participants were more comfortable interpreting their voices as relationships and not as the sign of a violated mind [… Given this] harsh, violent voices so common in the West may not be an inevitable feature of schizophrenia." These claims chime precisely with the many persons labeled as having schizophrenia who have joined self-advocacy organizations argue. They not only find seeing their condition as a difference to be technically accurate, but also, significantly, simply much more helpful than the "treatments" they received as service-users within institutional psychiatry (Farber 2012).

Bearing this in mind, and although these various authors do not write from a specifically neurodiversity paradigm perspective, their key insight—that the distress experienced by people with schizophrenia may stem from social and relational rather than internal factors—gives some support for the neurodiversity paradigm as applied to schizophrenia. In other words, given that schizophrenic persons can live, in the right society, in such a way that they are not distressed to a clinically significant level—and this is before we even get to the strengths associated with schizophrenic cognition (Badcock 2009)—then it is far from clear that the condition should qualify as a harmful dysfunction. On the contrary, dominant ideals of selfhood seemingly exclude and harm large numbers of people who fall outside them, including schizophrenics. This provides further reason to trace their distress back to subtle yet pervasive social and ideological factors as much as to inherent pathology. Moreover, given that integrating one's schizophrenia in an authentic self-narration would, presumably, be aided by the neurodiversity paradigm's more affirmative stance toward how individuals relate to their conditions—meaning that there is some initial reason to think that this may ultimately be more pragmatic in respect to fostering schizophrenic flourishing—something that the pathology paradigm proponent should, as was mentioned earlier, consider worth taking into account.

Concluding remarks

What can we conclude from this? I certainly do not take myself to have proven, once and for all, the all cases of multiply disabled autism or schizophrenia can adequately be accounted for on the neurodiversity paradigm (although I do believe that this might be justified in the future). Rather, I mean to disrupt the assumption that such cases are obviously problematic for the neurodiversity paradigm. For, what it seems is that whether one paradigm is ultimately more right than the other will come down not to scientific research, but rather to (sometimes arbitrary or unjustified) underlying philosophical beliefs regarding, say, what it means to be a self, what it means to flourish as a human, to what extent species-standard norms should delineate our judgments in this regard, and so on.

In fact, using the examples of multiply disabled autism and schizophrenia in order to criticize the neurodiversity paradigm actually begs the question: for it needs to presuppose precisely the various kinds of conceptual commitments and normative judgments that the pathology paradigm is bound up with in order to reach the conclusion that these are, indeed, best framed as medical pathologies. Given this, if the pathology paradigm proponent wants to justifiably argue that

they are in fact problematic for neurodiversity paradigm proponents, they will need to provide a deeper and more nuanced kind of argument—one that justifies conceptions of what it means to be a self, the nature of human flourishing, and so on. Equally, of course, the neurodiversity paradigm proponents may also beg the question: by assuming that all (or a much wider range) ways of being are simply natural and legitimate, it may lead to overlooking some problems that really are internal to various individuals or groups.

One conclusion, then, might be as follows: taking the neurodiversity paradigm, or at least the views of its more nuanced proponents, seriously, means confronting both how deeply it challenges standard psychiatric thinking and practice, and in turn how intertwined psychiatric theory is with broader normative and ideologically laden judgments. And if this is the case, this challenges us to reexamine all categories and concepts that currently fall under the pathology paradigm, lest it turn out that we have wrongly medicalized suffering in any strand of human life that might better be accounted for in terms of social marginalization and oppression.

Acknowledgments

I am grateful to the Shirley Foundation for funding my research on autism and neurodiversity.

Notes

1 In the leading journal, *Psychiatry, Psychology, and Philosophy*, for example, there are no papers dedicated to the subject of neurodiversity, and almost all papers are firmly rooted in the very model that the neurodiversity movement challenges. Similarly, in the leading book series in the discipline, *International Perspectives in Philosophy and Psychiatry*, published by Oxford University Press, I have only found one brief discussion of the movement, despite the subject of neurodiversity being increasingly studied in other disciplines such as disability studies, bioethics, and politics.

2 Although it is widely agreed that mental disorder refers to harmful dysfunction in relation to some norm or another, there is disagreement regarding how best to frame and measure this (Bolton 2008; Glover 2014). The most influential accounts come from Boorse and Wakefield. Boorse takes a bio-statistical approach, where we can discover the "normal" functions of humanity in relation to species-standard statistical norms. For example, it is clearly the case that the statistical majority of hearts pump blood, and so we can take this to be their "normal" function of the

heart. In turn, any deviation from this, if it is harmful at least, can be taken to be pathological. By contrast, Wakefield suggests an evolutionary norm, whereby any harmful deviation from the evolutionary function can be taken as pathological. On this conception, given that hearts evolved in order to pump blood, then any harmful deviation from this evolutionary norm is a harmful dysfunction. So for these pathology paradigm proponents, abnormality in the sense of harmful dysfunction can be identified in relation to bio-statistical or evolutionary norms.

3 Although, it might be replied here that since this condition stops *any* self whatsoever developing, rather than it being a matter of delineating the *kind* of self that develops, then Walker's understanding of the neurodiversity paradigm is compatible with seeing the condition as a medical pathology.

4 Autism could also be prevented, for example, if a prenatal test were developed that allowed the systematic abortion of autistic children. On this see Virginia Bovell (2015), who has developed the most nuanced framework for understanding what this entails. On her account, argued in great detail, attempts to cure or prevent autism lack moral legitimacy.

5 As it turns out, though, proponents of the (now less active) "mad pride" movement have argued precisely for a reevaluation of schizophrenia, in a manner that largely overlaps with the (now more active) neurodiversity movement at least. In any case, rather than simply reiterating the arguments of mad pride proponents here, I will examine schizophrenia in relation to the neurodiversity paradigm specifically as outlined above.

References

American Psychiatric Association (2013), *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Arlington, VA: American Psychiatric Publishing.

Badcock, C. (2009), *The Imprinted Brain*, London: Jessica Kingsley Publishing.

Baron-Cohen, S. (2011), *Zero Degrees Empathy: A New Theory of Human Cruelty and Kindness*, London: Allen Lane.

Bolton, D. (2008), *What Is Mental Disorder? An Essay in Philosophy, Science, and Values*, Oxford: Oxford University Press.

Bovell, V. (2015) "Is the Prevention and/or Cure of Autism a Morally Legitimate Quest?' DPhil. University of Oxford.

Bridget, A. (2012), Autism and Self-Injurious Behaviours, *The Thinking Person's Guide to Autism*. URL= http://www.thinkingautismguide.com/2012/10/autism-and-self-injurious-behaviors.html Retrieved May 1, 2017.

Casanova, M. (2013), The Neurodiversity Movement: Good Intention Resting on a Shaky Scientific Foundation, *Cortical Chauvinism*. URL= https://corticalchauvinism.

com/2013/06/11/the-neurodiversity-argument-good-intentions-resting-on-a-shaky-scientific-foundation/ Retrieved July 30, 2016.

Casanova, M. (2015), The Neurodiversity Movement: A Lack of Trust, *Cortical Chauvinism*. URL= https://corticalchauvinism.com/2015/01/05/the-neurodiversity-movement-lack-of-trust/ Retrieved July 30, 2016.

Chapman, R. J. (2016), Autism Isn't Just a Medical Diagnosis—It's a Political Identity, *The Establishment*. URL= https://theestablishment.co/autism-isnt-just-a-medical-diagnosis-it-s-a-political-identity-178137688bd5 Retrieved August 20, 2016.

Farber, S. (2012), *The Spiritual Gift of Madness, The Failure of Psychiatry and the Rise of the Mad Pride Movement*, Rochester, VT: Inner Traditions.

Frith, U. (2008), *Autism: A Very Short Introduction*, Oxford: Oxford University Press.

Glover, J. (2014), *Alien Landscapes? Interpreting Disordered Minds*, London: Harvard University Press.

Jaarsma, P. and Welin, S. (2012), Autism as a Natural Human Variation: Reflections on the Claims of the Neurodiversity Movement, *Health Care Analysis*, 29(1): 20–30.

Kuhn, T. (1962), *The Structure of Scientific Revolutions*, Chicago, IL: University of Chicago Press.

Luhrmann, T. M. R., Padmavati, H., and Tharoor, A. O. (2014), Differences in Voice-Hearing Experiences of People with Psychosis in the USA, India and Ghana: Interview-Based Study, *The British Journal of Psychiatry*, 206:41–44.

Merriam, G. (2010), Rehabilitating Aristotle: A Virtue Ethics Approach to Disability and Human Flourishing. In *Philosophical Reflections on Disability*. C. D. Ralston and J. Ho, eds. Dordrecht and New York: Springer Verlag.

Mottron, M. (2011), Changing Perceptions: The Power of Autism, *Nature*, 479:33–35.

Sass, L. A. (2004), Negative Symptoms, Commonsense, and Cultural Disembedding in the Modern Age. In *Schizophrenia, Culture, and Subjectivity*. J. H. Jenkins and R. J. Barrett, eds. Cambridge: Cambridge University Press.

Scotch, R. and Schreiner, K. (1997), Disability as Human Variation: Implications for Policy, *The Annals of the American Academy of Political and Social Science*, 549:148–159.

Sinclair, J. ([1993] 2012), Don't Mourn for Us, *Autonomy: The Critical Journal for Interdisciplinary Autism Studies*, 1.

Timimi, S., McCabe, B., and Gardner, N. (2010), *The Myth of Autism: Medicalising Men's and Boy's Social and Emotional Competence*, London: Palgrave Macmillan.

Walker, N. (2013), Throw Away the Master's Tools: Liberating Ourselves from the Pathology Paradigm, *Neurocosmopolitanism*. URL= http://neurocosmopolitanism.com/throw-away-the-masters-tools-liberating-ourselves-from-the-pathology-paradigm/ Retrieved June 14, 2016.

Walker, N. (2014), Neurodiversity: Some Basic Terms and Definitions, *Neurocosmopolitanism*. URL= http://neurocosmopolitanism.com/neurodiversity-some-basic-terms-definitions/ Retrieved June 17, 2016.

19

Managing Individuals and Populations through Psychiatric Classification

Devonya Havis and Melissa Mosko

Introduction

In this chapter, we argue that psychiatry works to manage individual bodies and populations. As a practice that creates specialized knowledge while simultaneously shaping social relations, we argue that psychiatry operates through what Michel Foucault describes as disciplinary power and biopower. In this respect, power is exercised through the creation of norms for behavior that work to force individuals within those norms, correcting abnormal behavior, and reinforcing the boundaries of the normal. In the modern state, technologies of power are used to create, disseminate, and reinforce social norms to control individuals, making them docile in relation to the extant power structure. Biopower is exercised through the classification of individuals into groups, in view of controlling the birth, life, and death of what are classified as desirable and undesirable populations. By regulating or minimizing undesirable populations, states can create "safe" environments for desirable populations to thrive in the name of social hygiene. This chapter analyzes how mechanisms of disciplinary and biopower operate through classification by looking at three cases: state-enforced sterilization of poor women and women of color, state-sanctioned lobotomies of women in the twentieth century, and state attempts to neutralize perceived social threats during the nineteenth and twentieth centuries by anticipating "dangerous" individuals and populations. Contemporary state use of crime predictive technology plays a similar anticipatory role by relying upon algorithms to assess potential risk for the commission of future crimes. These case studies demonstrate how psychiatric practices engage both disciplinary power and biopower.

Foucault on biopower, disciplinary power, dangerous individuals, and social hygiene

Michel Foucault, in lecture eleven of Society Must Be Defended (March 17, 1976), argues that the exercise of power in modern states depends upon mechanisms that are familiar but that, nevertheless, are not obvious. In fact, the very conditions of what is defined as "normal," legitimate, acceptable are always already conditioned by power relations—even when coercion does not seem evident. Foucault, in this lecture and the Abnormal lectures, links the everyday operations of the modern state to racism. The notion of "racism" that he employs includes conventional understandings of race but also expands understandings of "racism" to include a new formation described as "racism against the abnormal." For Foucault, psychiatry's historical function is tied to the generation of a "technology of the abnormal" which prompts social interventions and controls aimed at protecting society from perceived danger. This "technology of power" operates on individual bodies and populations both independently and in correspondence with what Foucault describes as traditional Western "ethnic racism." He writes:

> The racism that psychiatry gave birth to … is racism against the abnormal, against individuals who, as carriers of a condition, a stigma, or any defect whatsoever may … transmit to their heirs the unpredictable consequences of evil, or rather of the non-normal, that they carry within them. It is a racism whose function is not so much the prejudice or defense of one group against another as the detection of all those within a group who may be carriers of a danger to [society]. It is an internal racism that permits the screening of every individual within a given society. (1974: 316–317)

Even while race-based themes may not be explicit, Foucault's power analysis exposes how racialized themes that operate through normalization become embedded in the very mechanisms by which the state operates. As he notes, "the theme of race does not disappear, it … become[s] part of something very different, namely state racism" (1976: 239). The roadmap given in this 1976 lecture exposes the complex and often invisible set of mechanisms that condition what we take as given or reasonable. This is a form of normalization that makes it difficult for most of us to understand that what we take as "normal" is often a coercive effect of power relations. Foucault's analysis demonstrates the ways in which seemingly neutral mechanisms of state power function as anchorage points for state-supported racism.

While we tend to categorize all forms of racism as purely individual, Foucault makes evident that, racism is, in fact, a feature of the modern state. Racism is linked not only to individual bodies but also to populations within the state. Simplistic notions of racism that describe it as a matter of individual hatred or personal preference limit the capacity to understand how race is implicated in the nexus of power relations prevalent in modern states. These "everyday" notions of racism limit the extent to which racialized thinking can be connected not only to individual practices but also to institutional practices and laws. Racism establishes an interconnection between forces that govern individuals and populations. In Foucault's language, it invokes the simultaneous operation of two distinct forms of power: disciplinary, which is focused on managing individual bodies, and biopower, which is concerned with regulating populations.

Disciplinary power creates norms for individual behavior and transforms individuals into "docile bodies" who operate according to those norms. The mechanisms by which we train school children to be "good" pupils and "good" citizens provide a useful illustration of how disciplinary power functions through processes of normalization. Correct training is aimed at managing individual bodies. It is not simply a matter of punishment, but also entails rewards for desired behavior. Children who do not meet the norms are exposed to additional discipline and training rather than simply being punished. The remedy for failing to meet norms is more discipline.

Biopower, on the other hand, concerns itself not with individual bodies but with entire populations. In Foucault's words, biopower is "addressed to a multitude of men, not to the extent that they are nothing more than their individual bodies, but to the extent that they form, on the contrary, a global mass that is affected by overall processes, characteristic of birth, death, production, illness and so on" (1975: 243). With biopower, one does not see individuals as individual but rather as an entire population that is known based upon its collective statistics which concern birth, death, health, and so on. In this respect, an individual is merely a particular example of a larger group—a population—which has certain characteristics that have been objectively assigned based upon the group's collective data. Such data gives states and institutions information about the traits that characterize each population. In this way, states can promote healthy characteristics and healthy populations by eliminating those elements and populations that sap the state's strength. By minimizing undesirable irregularities or regulating populations, the state invokes biopower to create safe environments. As a result, there is a relationship between the flourishing of populations with desirable characteristics and the

withering of populations with undesirable characteristics. In Foucault's words, populations become a "political problem." It, therefore, becomes necessary to take control over life to let desirable populations live and let undesirable populations die. It is only through the death of undesirables that the desired elements can flourish. This is a form of public hygiene, aimed at reducing those elements that "sap a population's strength," the elements that "waste energy and cost money" (1976: 244–249).

According to Foucault, all modern states have a system of population management based on a hierarchy of certain "human species." Biopower generates a framework through which one comes to understand that some human sub-species are more valuable than others because of the positive traits that regularly occur within those populations. Meanwhile, other populations of human sub-species are considered non-normative and observed to regularly exhibit undesirable traits which must be excised. Hence, the state must engage in public hygiene because its very existence becomes linked to maximizing its resources by preserving the upper sub-species, by letting them live and flourish, while allowing the lower sub-species to wither and die. Public hygiene halts the threat of creeping degeneration from the lower sub-species by limiting "pollution" from criminals. It promotes the flourishing of the dominant sub-species, the right-living group, by rewarding desired behaviors and populations. Even though biopower operates at a different level from disciplinary power, it infiltrates and embeds itself by means of existing disciplinary mechanisms.

Psychiatry as a social institution

Psychiatry operates as disciplinary power and biopower because it is not merely an academic discipline or medical practice, but a social institution. Psychiatry is a social institution that is embedded and intertwined with other social institutions, and therefore contains and reinforces social assumptions. The concept of social institution used here is a holistic one, as opposed to an atomistic one (Miller 2009: 27). A holistic definition of social institutions highlights the inter-relatedness of various social institutions, and those institutions' contribution to broader and more complete social systems. Psychiatry is a social institution, for instance, because it is interrelated with academic and scientific research through the development and classification of disease and disorder criteria—such as the *Diagnostic Statistical Manual of Mental Disorders* (*DSM*) created

by the American Psychiatric Association (APA). It is interrelated with the legal and carceral system because psychiatric diagnoses and diseases are used as evidence of a person's reliability or culpability, and it is it interrelated with social institutions of gender and race, as these social groupings have been pathologized over time (c.f. Tekin and Mosko 2015). Psychiatry is a social institution that contributes to a more complete social system like American society in that it is a major social institution through which we live out our lives and which distributes rights and responsibilities according to our basic principles of justice (Rawls 1971). For instance, a person who is found to be a danger to themselves or to the public because of psychosis or bipolar disorder can be involuntarily hospitalized as a measure of social protection (Sederer 2013). Insofar as the academic and medical field of psychiatry develops and refines the symptoms and signs used to diagnose an individual, and such a diagnosis can be used by legal and civil authorities to limit a person's freedom, psychiatry is a social institution that contributes to American society.

In addition, the practice and social institution of psychiatry contains and reinforces social assumptions and unavoidably carries bias (Caplan and Wilson 1990). In the clinical setting, bias may arise from the clinician's individual psychology, including cognitive biases, her professional identity and practices, and contextual factors of her practice (Poland and Caplan 2004). Relevant contextual factors that create bias are the increasing demands placed on clinicians with decreasing time spent with clients, the influence of the pharmaceutical industry on research and development of diagnostic classification and pharmacotherapy-based solutions to mental disorder and disease. Beneath all of this is a Western cultural belief that there is a solution for all problems, and that these solutions can and will come quickly in the form of science, technology, and medicine (Poland and Caplan 2004). Clinicians themselves are persons, and all persons have social and cognitive biases that will unavoidably make their way into clinical practice. The interrelatedness of psychiatry with other social practices, like the lobbying influence of the pharmaceutical industry, and social beliefs about solutions to psychiatric problems listed above, render psychiatry—like every other social practice—susceptible to social bias and, ultimately, the state's use of power.

This description of social institution falls under what Foucault describes as *discourse*, or ways of speaking and meaning-making that constitute subjects, the world, and knowledge about both (Foucault 1969). Foucault himself argues that madness is socially constructed through power relations, variable and dependent upon the society that seeks to define madness (Foucault 1961). For example,

defiance in women was considered a symptom of mental illness in the late nineteenth century, grounds for diagnosing a woman with hysteria. In present times, defiance can be seen as a mark of individual psychological strength or ability to resist undue pressure and influence.[1] In this way, a society, through various state mechanisms and practices, produces "the mad" by classifying them as outside of or resistant to social norms.

Women, madness, and technologies of feminization

When women's lives intersect with psychiatric practice, Foucault's argument that psychiatry has generated a "technology of the abnormal" by creating social interventions on both individual bodies and whole populations comes into view. Two cases in the early and mid-twentieth century exemplify the development of technologies[2] used to control women's bodies in order to normalize them and make them more docile. However, more than just control of *women,* we also see the control of women's bodies as a technique of social hygiene, limiting women's reproductive capacities to increase the genetic fitness of the overall population. Foucault argues that disciplinary power is exercised over bodies themselves, creating docile bodies which operate according to the social norms defined by a given society. This appears in the use of psycho-surgical lobotomy on women to make them docile, infantile, and "feminine." As we show below, the technique of lobotomy to create docile bodies stands in stark contrast to the budding women's movements of the 1920s and 1930s which saw women's participation in social and political life increase, as well as an increase in women's demands for political representation.

Psychiatric technique as a form of control over women's bodies didn't end with the aim of creating docile women, however. Women's bodies became a political tool through programs of forced sterilization in the twentieth century as well. As a form of what Foucault identifies as biopower, forced sterilization practices used population statistics and the mechanisms of psychiatry to control who could reproduce in the name of social hygiene. Carrie Buck, whose case would set the legal precedent for forced sterilization of the feeble-minded, was argued to be genetically unfit to reproduce, as her mother had been institutionalized for feeble-mindedness, immorality, and prostitution. Mental disorders, sexual promiscuity, and criminality were thought—with the help of diagnostic tools of psychiatry—to be genetically inheritable traits, and thus it was in the best interest of the whole society to limit the "unfit" from reproducing.

Forced sterilization and population control

In twentieth-century United States, programs of forced sterilization utilized psychiatric classification as a form of biopower and disciplinary power. Between 1897 and 1977, thirty-three states in the United States had programs for the involuntary sterilization of individuals with mental illnesses or disorders. Though there were challenges to these practices, the US Supreme Court ruled in the infamous *Buck v. Bell* case that the sterilization of persons with mental disorders did not violate the US Constitution. At issue in *Buck v. Bell* was whether Dr. Albert Sidney Priddy had the authority to involuntarily sterilize Carrie Buck on the grounds that she was feeble-minded, possessing the intelligence of a nine year-old. Carrie's mother, it was claimed, had the intelligence of an eight year-old, and a reputation for "immorality" and prostitution. As it would turn out, Carrie Buck was not feeble-minded (Lombardo 2008). The prevailing belief at the time was that poor children, like Carrie, would be better off if they were removed from their homes to grow up in middle-class households. Carrie was abused and exploited in her foster home, and when she was impregnated by the nephew of her foster mother, her foster parents sent her to the Virginia State Colony for Epileptics and the Feebleminded. She became a test case for the then-new eugenics sterilization law in Virginia. The law allowed for the sterilization of those "afflicted with hereditary forms of insanity that are recurrent, idiocy, imbecility, feeble-mindedness, or epilepsy" (Landman 1932: 84). Because Carrie's mother was also institutionalized at the Colony, it was argued that Carrie's ability to reproduce posed a genetic threat to society in that her children would also be feeble-minded and a burden on or threat to society, triggering Justice Oliver Wendell Holmes, Jr. to say in the majority opinion of the court, "three generations of imbeciles are enough." In all, between 60,000 and 70,000 sterilizations were performed in the United States as a result of *Buck v. Bell*, most often on poor, white women.[3]

Eugenics movements were common at the time of *Buck v. Bell*. The 1924 Immigration Act limited the number of immigrants allowed to enter the United States, severely restricting the immigration of Southern and Eastern Europeans, specifically Jews and Italians, and completely restricting the immigration of Africans and Asians to the United States. Proponents of the 1924 Immigration Act often cited eugenics in support of it, claiming that the strength of the United States rested in part in the genetic makeup of its population (Baynton 2014). Prior to this, eugenics research and movements received significant funding from the foundations of the Carnegie, Kellogg, and Rockefeller corporations. In

1911, leading biologist Charles Davenport founded the Eugenics Research Office (ERO), which would go on to collect and analyze family pedigrees. The ERO forwarded the view that some persons were a threat to the general population, and thus genetically unfit for society. They found that the "unfit" usually came from economically poor families and backgrounds. Women's rights activist Margaret Sanger, foundress of the organization that would become Planned Parenthood, argued for birth control and family planning on similar grounds as eugenicists. She argued that those who were unfit to raise children or who had acquired undesirable traits that could spread to their offspring ought to limit their own reproductive ability (Sanger 1921: 5). Until the 1942 ruling in *Skinner v. Oklahoma* which severely restricted the practice, many states used sterilization as a punitive measure, forcibly sterilizing individuals who were convicted of crimes involving moral turpitude. While it did not overturn *Buck v. Bell*, and did not entirely eliminate punitive sterilization, *Skinner* drew attention to the often capricious uses of sterilization, and justices recommended strict scrutiny of all eugenics laws. It was around this time that Americans' opinions on eugenics were shifting, largely due to the eugenics programs of Nazi Germany that fed the Holocaust (Sullivan 2002).

The two main targets of forced sterilization in the American eugenics movement were criminality and mental illness or mental defect. Both were thought, at the time, to be genetically inheritable traits, and both were also believed to be tied to economic and social class. Doctors in psychiatric institutions and in prisons used psychiatric classifications to make the case that individuals were proper candidates for sterilization, and that sterilization of "unfit" individuals was necessary for the social good. Deploying psychiatric classification and emboldened by the ideals of progressivism, eugenicists exercised power over the bodies of individuals to increase their usefulness to society, or at least prevent them from producing offspring believed to become a drain on social resources. Linking feeble-mindedness and criminality to the genetic disposition of impoverished populations, eugenicists effectively managed poor populations through forced sterilization or voluntary forms of birth control and family planning, as in the case of Margaret Sanger's advocacy.

While the most widely known cases of forced sterilization occurred in the mid-twentieth century, as recently as 2010, practices of forced sterilization have required legal and political intervention. Between 2006 and 2010, nearly 150 females in California's prison facilities underwent forced sterilization, prompting the state legislature to ban sterilization of female inmates except for explicit medical reasons in 2014. It is estimated that the State of California paid

nearly $150,000 for tubal ligations as part of its sterilization protocol. A former obstetrician and gynecologist for Valley State Prison in California said he offered the procedure as a service to women, and that the cost to the State over the ten-year period is not high, "compared to what you save in welfare paying for these unwanted children—as [the prisoners] procreated more" (NPR 2013).

Twentieth- and twenty-first century programs of forced sterilization are prime examples of both disciplinary power and biopower. Disciplinary power establishes norms for individual behavior and creates individuals to be docile. As in the case of Carrie Bell, removed from her home and placed in the care and servitude of a middle-class family, the individual stands as a threat to the social order through some bodily (or psychiatric) affliction. Bell's sterilization and the sterilization of thousands of other women allowed them to better operate according to social norms and to better fulfill their social function. The practices of forced sterilization upon those diagnosed as insane or those charged with being criminal exhibit the use of psychiatric classification as a form of biopower. Classification groups otherwise distinct and unique individuals together in order to control these newly formed populations for the sake of social hygiene. As noted above, populations become a political problem, and thus it is in the interest of a whole society to let desirable populations live and let undesirable populations die. Sterilizing those thought to be genetically deficient through insanity or criminality effectively lets undesirable populations die so that desirable populations can flourish.

Lobotomy and the legacy of hysteria

One of the chief victories of psychiatry in the mid-twentieth century was the development of psychosurgical procedures such as the lobotomy, which stood beside insulin-shock therapy and electroconvulsive shock therapy (ECT) as the greatest medical advancements in psychiatry in generations. Each of these was developed to cure or abate symptoms and signs of psychiatric illness that could not otherwise be controlled or managed except through institutionalization. The modern form of lobotomy was developed by Portuguese neurologist Antonio Moniz in 1936 as a frontal lobectomy which severed the cellular connections between the frontal lobes of the brain, believed to be mal-formed in patients suffering from psychiatric illness. In the United States, Walter Freeman and James Watts worked to develop the technique, notably pioneering the transorbital lobotomy in 1945. Given the popular success of Moniz's frontal lobectomies, Freeman and Watts aimed to simplify the procedure and make safer

to perform in a variety of medical situations. The frontal lobotomy required drilling holes in the skull, and with such high-morbidity risk, required that the procedure be performed in well-equipped surgical suites. Freeman adapted the procedure so that the surgeon could reach the patient's brain through the orbital socket, alleviating the need to drill into the skull. For Freeman, this meant that the transorbital lobotomy could be transported from the secure surgical suite into the state mental hospitals where it was needed, and which lacked surgical facilities.

In the 1940s and 1950s, at least 20,000 lobotomies were performed in the United States alone, and innumerable thousands more around the world. Proponents argued that when successful, lobotomy normalized patients' behavior (Walker 1993: 25). In the worst of cases, lobotomy patients either did not survive the operation, died as a result of it, suffered memory loss, experienced loss of bodily function, or developed dementia. It is now well-documented that the majority of lobotomies during this time were performed on women, who were also the most likely candidates for ECT. There are detailed, gendered reasons for performing more of these procedures on women, namely, that the dulling effects of lobotomy and ECT were appropriate side effects for women who would then become more docile and better able to fulfill domestic duties (Johnson 2014: 45; Walker 1993: 28).

Lobotomy thus had both therapeutic *and* social value: it feminized women. In the 1920s and 1930s, women sought out new forms of self-expression and social engagement that were characteristically un-feminine for the time. Psychiatric practice thus "whether purposely or inadvertently, used their treatments to adapt deviant women to traditional gender roles" (Walker 1993: 23). For example, in 1946 a young Michigan woman known as "Case 1000" was lobotomized not because she was mentally ill, but because she was known to violate laws and social norms. Her case was widely publicized, the press often noting not mental illness or criminality, but the fact that she liked to smoke cigars, a very un-feminine activity (Johnson 2014: 44). Lobotomies were also performed on aggressively sexual women and women with abnormal sexual desires. The woman described in Case 1000 is also said to have suffered from aggressive sexuality, but there were many married women who underwent lobotomies because they had sex outside of their marriages, or who were just unhappy in them. *Jet* magazine reported on the case of a woman who was sent for surgery because she was irritable and would throw things at her husband. So, in addition to sexual control of unmarried women, psychosurgery was used to control women within marriages, to restore the "harmony" within heterosexual marriages (Johnson 2014: 58).

While not as widely discussed in the literature, lobotomy was hailed successful for its ability to normalize gender for men as well. It was used on men who wore women's clothes, and a 1953 *Los Angeles Times* story describes the case of a businessman who was dissatisfied in his marriage and attempted suicide by firing six bullets into his head. He survived, and a lobotomy was performed. Freeman, his doctor, noted that upon recovery, the man divorced his wife, and then went on to professional and financial success. Given that professional and financial success are hallmarks of masculinity (Johnson 2014: 58), this man's lobotomy helped him fit into social norms of masculinity.

Psychiatric practice intersected with gender normalization a half-century before the rise of the lobotomy, in the scientific study of hysteria. While the history of hysteria is rife with myth and legend (King 1993), in the late nineteenth century Jean-Martin Charcot embarked on the first *scientific* study and classification of hysteria. Unlike other mental illnesses, only women were diagnosed with hysteria, based on showing signs of motor paralysis, sensory loss, convulsions, and/or amnesia. While Charcot spent his career taxonomizing symptoms and presentations of hysteria as an observer, those who followed after him, like Sigmund Freud, sought the causes of hysteria. Through lengthy processes of interviewing patients, Freud and others discovered that female hysterics—often suffering from lives of violence, exploitation, and rape—all shared experiences of early sexual trauma. While Freud would go on to deny the reality of female hysterics' experiences in favor of repressed fantasy and desire, the connection between hysteria and women's sexuality remained in the popular imagination (Gilman et al. 1993). For his part, Charcot would demonstrate hysteric fits in patients to large audiences, popularizing images of hysteria in both literature and politics. Hysteria took on a life of its own in the popular imagination.

As one author notes, psychiatrists didn't set out to study women, but women and the mystery of hysteria served an important political purpose for scientific psychiatry (Herman 1992: 10). With the decline of monarchy and the rise of republican secularism in France, replacing the "mystery" of hysteria with clinical observation and taxonomy represented the success of the scientist's role in the new secular society. In claiming women and hysteria for science, psychiatrists objectified the experiences of women in their pursuit of replacing religious mythology with scientific understanding.

The history of hysteria in the late nineteenth century is instructive for putting the use of lobotomies on women in political perspective. Women became useful to secular, scientific society as objects of study. What had been for centuries

a "mysterious malady" suffered by women was put under clinical observation. Scientific description and classification of hysteria and thus, women, allowed male scientists to claim a victory in the nineteenth-century struggle between the democratic ideology accompanying scientific secularism and the authoritarian ideology that would maintain society's religiosity. In addition, the alignment between hysteria and sexual trauma—which would continue well into the twentieth century in the development of diagnostic criteria for post-traumatic stress disorder—created an ideological synthesis of women, mental illness, and sex and sexuality (Herman 1992). This appears in the use of lobotomies on women in the mid-twentieth century. Lobotomy, as well as ECT, was used to "normalize" women into performing appropriate gender roles, and as is cited, was often used to control women's sexuality and confine it to heterosexual marriage. Lobotomy was used as an exercise of disciplinary power as it was deployed to control women's bodies creating them as docile, domesticated, child-like persons who could be more easily controlled by society, their caregivers, or husbands. This was socially necessary as women began challenging the confines of gendered norms that restricted their social, economic, and political lives in the 1920s and 1930s. In fact, resistance to treatment was categorized as a sign that one needed treatment, effectively pathologizing women's resistance to control over their bodies and minds. We see in both the cases of hysteria and lobotomy that social hygiene is used as a rationale for controlling women as a population.

Constructing "the dangerous individual" and crime predictive technology

Foucault's insistence that "legal justice has at least as much to do with criminals as with crime" (2013: 128) is instructive and provides a way to understand mechanisms of classification aimed at identifying and managing populations who are characterized as potential dangers because they threaten society from within. The heightened focus on developing a calculus for detecting the signs of potential criminality in individuals and populations constitutes what Foucault calls the "psychiatrization of criminal danger." This impetus to "register" criminals who would, under prevailing conditions, escape notice as dangerous. It is at this historical juncture that psychiatry becomes formally linked with the problem of identifying the "dangerous individual." This concern emerges in the nineteenth century and inscribes psychiatry into judicial processes. Psychiatry and criminology undertake the "new" problem—making intelligible, and

thereby visible, those criminals who commit crimes that appear to be "without material motive." Hence, psychiatry and criminology must develop a register that signals a dangerous element even when there are no visible marks to signal danger. The looming threat indicates the necessity to introduce psychiatry into the legal machinery to insure "public hygiene" (Foucault 2013: 130–134). The nineteenth-century focus on detecting criminal danger before it is manifest can be linked to contemporary predictive policing efforts that attempt to detect future crime before it occurs. In effect, predictive policing draws upon the historic connection between psychiatry and criminology that Foucault explicates in his analysis of the problem of "the dangerous individual." The historic mechanisms for registering "dangerous individuals" as well as the concerns about these individuals and populations support use of predictive policing technology and undergird its operation.

The concept of "the dangerous individual" emerges alongside concerns about crimes that were shocking and unsettling precisely because they were accompanied by no preceding "visible signs" that would demarcate the one who committed the crime as dangerous. Since such crimes were not accompanied by the usual signs of insanity and were, according to the "experts," seen as "crimes against nature, against those laws which are perceived to be inscribed directly on the human heart" (Foucault 2013: 131), psychiatrists sought to enter the legal machinery by developing an anticipatory technology that would neutralize social threat by "treating" danger and "curing" social illness. This medicalized model, that characterized crime as illness, animated a technology of punishment focused on the individual rather than the crime.

In effect, those who committed such crimes were monstrous for the experts because the lack of prior indications made the crimes, in their estimation, unintelligible. Pathological crime was marked by "the great and monstrous murder without reason, without preliminaries, the sudden eruption of the unnatural in nature" (Foucault 2013: 132). Such crimes prompted the need for a criminal psychology that could "function as … public hygiene." The aim, then, of medicalized psychiatry was to treat social danger—characterized as "danger for oneself, others, one's contemporaries, and also one's descendants through heredity" (Foucault 2013: 134). The idea then was a mechanism that would allow "the control of the dangerous in human behavior" (Foucault 2013: 135). The "dangerous" became categorized as a "criminality indicating a derangement with no other symptom than the crime itself." It was viewed as "a crime which is insanity and an insanity which is crime" (Foucault 2013: 132). To further elaborate, the dangerous individual is viewed as having an "insanity" and

embodying a criminality that "transgresses all the laws of nature and society" and that remains invisible until it explodes (Foucault 2013: 134–135). In effect, the individual's insanity is thought to be confirmed by the individual's crime and the crime is seen as evidence of the individual's insanity. The "danger" is marked by action that has "a maximum of consequences" with "a minimum of warning. The most effects and fewest signs" (Foucault 2013: 135). The crime is the event that signals the existence of a dangerous element and also signals a danger to the social body. Managing this "danger" means that "legal justice" must have as much to do with criminals as with crimes.

According to Foucault's analysis, the criminal consequently comes to be considered mad as indicated by the commission of crimes that ostensibly are "without reason" (Foucault 2013: 136). Such crimes were viewed as particularly problematic because administering criminal justice implied a "reasoned technology that would be applied" to an individual's punishment. Punishment was to bear on the criminal rather than the crime. A punitive mechanism functioning under this logic could not operate under conditions where the law authorized punishment but the courts could not develop a psychologically intelligible link between the act and its author (Foucault 2013: 138).

The dangerous individual, as threat, needed to be accounted for in discourses that integrated the acts of criminals into the global behavior of subjects. The "theme of the dangerous man," as an effect, gets inscribed into the institutions of psychiatry and justice as a mechanism for indexing behavior that was viewed as criminal within the context of people's general behavior (Foucault 2013: 139). This leads to the increasing focus on the dangerous individual in the nineteenth and twentieth centuries in which those labeled dangerous increasingly become the targets of punitive intervention … the anthropology of criminal man … and the theory of social defense. In this context, the intelligibility of the act with reference to the conduct, the character, the antecedents of the individual become the determinants of responsibility (Foucault 2013: 140).

Technology used to predict the potential for future crimes is a contemporary mechanism for registering "dangerous individuals" and populations. Such technology, it can be argued, seeks to manage individuals and populations in ways that determine danger before it is manifest. As such, the defense of society depends upon a public hygiene focused on identifying individuals and populations that threaten public order from within even before they act. Predictive policing is a means of neutralizing social threat using processes designed to register danger even when no visible signs appear. The ability to ferret out nonvisible signs indicating criminal danger rely upon psychiatry and

criminology which seek to register and "treat" what is considered dangerous in human behavior. This sensibility functions as the model for predictive policing technologies and, despite the proprietary nature of the technologies, is touted as the basis for its driving algorithms.

Generally called predictive policing, contemporary use of mathematical techniques designed to forecast or "predict" future criminals and future crime is considered cutting-edge technology. Literature promoting its use describes predictive policing as "the use of analytical techniques to identify promising targets for police intervention with the goal of preventing crime, solving past crimes, and identifying potential offenders and victims." The technologies available for use by law enforcement agencies fall into four categories with corresponding methods "for predicting offenders, for predicting perpetrators' identities, and methods for predicting victims of crime" (Perry et al. 2013).

Use of computer software and attendant technologies is viewed as innovative, more efficient, and objective by law enforcement entities. Law enforcement jurisdictions that face financial challenges are increasingly incorporating the use of predictive software models as a means of maximizing the ability of law enforcement to maintain effectiveness while operating with fewer dollars. It has been seen as a lower cost way to maximize crime prevention efforts when many jurisdictions must balance having access to fewer resources with the continuing need to manage populations. Conditions of resource scarcity and the desire to do more with less drive the focus on efficiency. The technologies are touted because of their ability to take large quantities of data and, in a short period of time, distill the information into categories differentiating the most desirable from the least desirable. Since math is combined with various forms of technology the evaluative process is presumed to allow systems to be more efficient and objective in producing results. Some of the models are used to predict crime in geographical areas. Rather than focusing on individuals, the software forecasts where crimes are most likely to occur allowing law enforcement to direct more resources to managing those areas (O'Neil 2016). Other predictive software focuses on individuals. In effect, such predictive technology uses mathematical models, or algorithms, to deploy classificatory mechanisms that determine the potentially dangerous by numerically scaling the likelihood that an individual will commit a crime in the future.

Within the discourse of psychiatry and criminology, risk assessment technology diagnoses and classifies "the dangerous individual" and does so on a vast scale. Scores generated using various types of future-crime-predictive technologies are used throughout the United States to inform decisions about

defendants at all stages of the judicial process from determining where to bolster police presence, who can be released without bail, bail amounts, length of sentence for a particular crime, to determinations about who must remain in prison, be granted probation or parole. Predictive policing is used throughout the world and at least twenty-four states in the United States use some form of the technology.

Much of the future-crime-predictive technology used in federal and state jurisdictions is owned by for-profit corporations this means that the algorithms driving classifications, or risk assessment scores, are proprietary and therefore secret. Moreover, many software models in use have not been tested for validity. Yet, they are routinely used to "score"—evaluate and classify—those who are arrested and incarcerated. A "Crime Reform" bill that has been introduced in US Congress could mandate the use of risk assessment technology in federal prisons. Former US Attorney General Eric Holder, in 2014, cautioned that using the technology could actually "exacerbate unwarranted and unjust disparities … already far too common in [the U.S.] criminal justice system and in our society" (Calabresi 2014). Within this context, Holder urged a comprehensive study evaluating the validity of risk assessment tools to determine their effectiveness and to determine if some groups are more likely to be wrongly labeled as higher risk. Like Holder, many have expressed concern about the use of such technology because there are currently no comprehensive data exploring whether predictive software produces scores that could be or that are racially biased (Angwin and Larson 2016).[4]

Supporters of risk assessment technologies argue that their processes of scoring will be an aid in reducing incarceration rates. As such, forecasting tools have seen an increased demand associated with more recent attempts to reverse mass incarceration. In fact, the for-profit company that produces one of the most widely used predictive software models celebrates its tool's ability to determine "criminogenic needs" by developing a "quantitative taxonomy" that measures such personality traits as intelligence, introversion, extroversion, while including major theories about the causes of crime in its scoring process (Angwin and Larson 2016).

Historically criminologists and criminal psychologists have taken the lead in developing criteria used to predict danger. Their charge has been to detect which individuals and populations might be the carriers of danger. With future-crime-predictive technologies, machines can now be enlisted to serve as "technicians" of the social body addressing the pressing problems of the urban streets, demographics, and industrial labor—those areas that have historically

been monitored as incubators of danger. The role of ensuring public hygiene, historically the purview of criminal psychiatry, can be automated through a contemporary machinery that functions to defend against the dangers inherent in the social body itself. In this respect, predictive policing invokes disciplinary power and biopower in its operation.

By focusing on the results that depend upon numerical data, these predictive technologies operate with the presumption of objectivity. What becomes obscured is that the mathematical applications driving results generated by crime-predictive-technology rely upon choices about how information will be correlated and what types of information will be emphasized as proxies for missing data. The consequences are that these seemingly objective tools obscure biases that are inadvertently imported into the models that drive results. As such, critics argue that the opacity and secrecy concerning the algorithms powering the technology disproportionately punish anomalies or exceptions—the poor and oppressed are disproportionately marked by such technology as potential dangers in need of management.

In an analysis of one of the most widely used predictive technologies in the United States, the study found that significant racial disparities emerged when using the software to predict who would reoffend. Errors in assessments for blacks and whites were made at roughly the same rates but the results encompassed a different type of error in forecasting. Black defendants were twice as likely as white defendants to be labeled as high risk for committing future crimes while whites were more often mislabeled as low risk. The blacks labeled as higher risk, over time, were actually less likely to actually reoffend while whites labeled as low risk actually went on to reoffend. The analysis evaluated assessments generated by the technology in Broward County, Florida and found that only 20 percent of those predicted to commit violent crimes actually did so. Even when isolating the effects of race from criminal history and recidivism, investigators found that black defendants were still 77 percent more likely to be assigned scores that marked them as higher potential risks than white defendants. They were also more often assessed to be at higher risk to commit a future violent crime and as 45 percent more likely to commit a future crime of any type. Despite these classifications that were based on scores from risk assessment technology, blacks, as a group, tended not to go on to reoffend while whites labeled as lower risk tended to go on to commit other crimes at much higher rates (Angwin and Larson 2016; Larson and Angwin 2016). The analysis examined and contrasted predicted rates of repeated offense with actual rates of repeated offense over a two-year period. In looking at the distribution of risk assessment scores for predictions of repeated

offenses in general and repeated violent offenses in particular, black defendants were evenly spread across scores while white defendants had scores clustered in lower-risk categories. The analysis also explored whether the difference between a high score and a low score for potential repeat offenses was different for black defendants and white defendants. The findings were that the "hazard" scores performed differently among racial subgroups. White defendants scored as high risk were 3.61 times more likely to reoffend than white defendants scored as low risk. This was contrasted with black defendants who, when scored as high risk, were only 2.99 times as likely to reoffend as those black defendants rated as low risk. Overall the mathematical classificatory technology was "more likely to misclassify a black defendant as higher risk than a white defendant." When restricting the classification of high risk to include only high (and not medium) scores, black defendants were 16 percent more likely to be falsely labeled high risk compared with 5 percent of white defendants (Larson and Angwin 2016).

Conclusion

Socially vulnerable individuals and populations are made more vulnerable when subjected to observation, evaluation, and management through classificatory systems. Like other methods of diagnosis and evaluation, psychiatric practice works by means of processes that seek to classify individuals and populations based upon existing social norms. These norms do not simply exist as givens but are formed through social practices that often create and reinforce existing power dynamics, leaving those who are deemed undesirable at even greater disadvantage. Managing "danger" and "undesirables" through processes of classification works to minimize or eliminate any resistance undertaken by individuals. It also facilitates elimination of those populations that threaten the social structure by letting them die through the control of birth, death, and life. In the cases of women's experience with lobotomy and forced sterilization, classification worked to identify individuals who were genetic threats to the population in order to prevent them from reproducing, or rendered them docile and "normal" given the prevailing ideals of femininity at the time. In each case, women who were already vulnerable on account of their gender in patriarchal society, or who were poor, considered sexually promiscuous, or suffering from mental disorders were controlled to the point of docility or genetic eradication. Race[5] figures prominently in the cases highlighting "dangerous individuals" and those predicted to commit future crimes. Yet, considerations of how race has

historically played a role in determinations of "danger" are rarely considered. The view is that determinations of criminality and predictions of the possibility of engagement in future crimes deal with race only incidentally. However, race remains deeply embedded within United States and global social structures even though explicit references to race are camouflaged. As Angela Davis notes, "The fear of the 'criminal'—the racialized figure of the criminal—has come to represent the most menacing enemy of 'American society.'" This makes the "racialized fear of crime … the quintessential enemy against which the nation imagines its identity." Crime and the prison, as well as mechanisms of classification that operate within their logic, work to simultaneously create and conceal the ways race is operative in everyday understandings of crime, drugs, immigration, and welfare. As such, "The enemy within is far more dangerous than the enemy without, and a black enemy within is the most dangerous of all" (Davis 2010: 270–271).

From the Foucauldian perspective taken up in this chapter, psychiatric classification disproportionately subjects socially vulnerable individuals and groups to state control. As such, psychiatric classification often constitutes a technology that fosters "racism against the abnormal." Efforts to detect and register the carriers of "danger" promote use of technologies that view individuals and populations not merely as mad or potentially criminal. These individuals and populations get labeled as "degenerate." Within this classificatory system, it is not only the individual who suffers from an "illness." The individual, actually suffers from the condition of "abnormality" which must be managed through psychiatric technology that intervenes, controls, and minimizes the threat to society by managing behavior and biology. As Foucault notes, this different form of "racism," historically linked to psychiatry, emerges in the nineteenth and twentieth centuries to detect those who might transmit the "condition" to their heirs along with the unpredictable consequences of the "non-normal that they carry within them." Hence, the function of this type of "racism" is not prejudice or defense of one's group against another group. Its focus is to determine everyone within a state who may pose a danger to the state. For this reason it permits the "screening of every individual within a given society." Moreover, attempts to mark the "abnormal" are grafted onto and operate alongside gender, race, and class biases. This means that in the name of social defense, there is an intensification of the relationship between the "abnormal condition" and the perception of transgression (Foucault [1975] 2004: 315–324).

Similar psychiatric procedures and notions, according to Foucault, continue to be operative in contemporary practices. One can anticipate their increasing

deployment through machinery capable of performing mass-scale classifications. The use of mathematical models produces evaluations that disproportionately ensnare and control those who are socially vulnerable frequently exacerbating existing inequalities. The evaluative algorithms are typically employed to manage large numbers of people at low cost which means that those who are poor and who experience other forms of disadvantage are increasingly subjected to evaluations by machines while those who are more privileged have the benefit of human intervention. Those harmed by imperfections in the mathematical models are accepted as "collateral damage … deemed unworthy and expandable" (O'Neil 2016: 13).

Notes

1 See Nancy Potter's chapter in this volume for a discussion of defiance and psychiatric diagnosis.
2 By "technologies" here we use Foucault's broad concept of technologies and techniques as practical rationalities governed by a conscious aim. These can be what we colloquially refer to as "technology" but need not be. C.f., Foucault, M. (1988).
3 For a discussion of the estimated cases of sterilization, see Cohen (2016).
4 See also Electronic Privacy Information Center (2017).
5 Race here refers to the historical process by which certain bodies as seen as having a race (i.e., black, Asian, Latina/o, etc.). This history creates conditions under which racialized bodies are also seen as threatening, exotic, ugly, or abnormal. Black bodies have historically been associated with criminality and therefore danger. For further discussion see Havis (2014).

References

Angwin, J. and Larson, J. (2016), Machine Bias, *ProPublica*. URL= https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing Retrieved August 21, 2018.

Baynton, D. C. (2014), *Defectives in the Land: Disability and Immigration in the Age of Eugenics*, Chicago, IL: The University of Chicago Press.

Calabresi, M. (2014), Exclusive: Attorney General Eric Holder to Oppose Data-Driven Sentencing, *Time*. URL= http://time.com/3061893/holder-to-oppose-data-driven-sentencing/ Retrieved August 21, 2018.

Caplan, P. and Wilson, J. (1990), Assessing the Child Custody Assessors, *Reports of Family Law*, series three, Toronto: Thompson Reuters.

Cohen, A. (2016), *Imbeciles*, New York: Penguin Books.

Davis, A. (2010), Race and Criminalization. In *The House That Race Built: Original Essays by Toni Morrison*, Angela Y. *Davis*, Cornel West, *and* Others *On Black* Americans *and* Politics in *America Today*. W. Lubian, ed. New York: Knopf Doubleday.

Electronic Privacy Information Center, *EPIC—EPIC v. DOJ (Criminal Justice Algorithms)*. URL= https://epic.org/foia/doj/criminal-justice-algorithms/ Retrieved December 12, 2017.

Foucault, M. (1961), *Madness and Civilization: A History of Insanity in the Age of Reason*, R. Howard, trans. New York: Random House.

Foucault, M. (1969), *Archaeology of Knowledge*, A. M. Sheridan Smith, trans. New York: Routledge.

Foucault, M. ([1974] 2003), *Abormal: Lectures at the Collége de France 1974–1975*, G. Burchell, trans., New York: Picador.

Foucault, M. ([1975] 1977), *Discipline and Punish: The Birth of the Prison*, A. Sheridan, trans. New York: Vintage Books.

Foucault, M. (1988), Technologies of the Self, pp. 16–49. In *Technologies of the Self: A Seminar with Michel Foucault*. L. H. Martin, H. Gutman, and P. H. Hutton, eds. Amherst: University of Massachusetts Press.

Foucault, M. (2013), The Dangerous Individual. In *Politics, Philosophy, Culture: Interviews and Other Writings, 1977–1984*, New York: Routledge.

Foucault, M., Bertani, M. M., Fontana, A., Ewald, F., and Macey, D. ([1976] 2003), *Society Must Be Defended: Lectures at the Collège de France, 1975–76*, Oxford: Macmillan.

Foucault, M., Marchetti, V., and Salomoni, A. ([1975] 2004), *Abnormal: Lectures at the Collège de France, 1974–1975*, Oxford: Macmillan.

Gilman, S. L., King, H., Porter, R., Rousseau, G. S., and Showalter, E. (1993), *Hysteria beyond Freud*, Los Angeles: University of California Press.

Havis, D. N. (2014), "Seeing Black" through Michel Foucault's Eyes: "Stand Your Ground" Laws as an Anchorage Point for State-Sponsored Racism. In *Pursuing Trayvon Martin: Historical Contexts and Contemporary Manifestations of Racial Dynamics*. G. Yancy and J. Jones, eds. Lanham, MD: Lexington Books.

Herman, J. (1992), *Trauma and Recovery: The Aftermath of Abuse—From Domestic Abuse to Political Terror*, New York: Basic Books.

Johnson, J. (2014), *American Lobotomy: A Rhetorical History*, Ann Arbor: University of Michigan Press.

King, H. (1993), Once upon a Text: Hysteria from Hippocrates, pp. 3–90. In *Hysteria beyond Freud*. S. Gilman, H. King, H. Porter, G. S. Rousseau, and E. Showalter, eds. Oakland: University of California Press.

Landman, J. H. (1932), *Human Sterilization*, Oxford: Macmillan.

Larson, J. and Angwin, J. (2016), How We Analyzed the COMPAS Recidivism Algorithm, *ProPublica*. URL= https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm Retrieved August 21, 2018.

Lombardo, P. A. (2008), *Three Generations, No Imbeciles: Eugenics, the Supreme Court and Buck v. Bell*, Baltimore, MD: Johns Hopkins University Press.

Miller, S. (2009), *The Moral Foundations of Social Institutions: A Philosophical Study*, Cambridge: Cambridge University Press.

NPR (2013), *California's Prison Sterilizations Reportedly Echo Eugenics Era* (2013), [Radio Program] July 9, 2013, 15:06.

O'Neil, C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York: Crown.

Perry, W. L., McInnis, B., Price, C. C., Smith, S., and Hollywood, J. S. (2013), Predictive Policing, Product Page. URL= https://www.rand.org/content/dam/rand/pubs/research_reports/…/RAND_RR233.pdf Retrieved August 21, 2018.

Poland, J. and Caplan, P. (2004), The Deep Structure of Bias in Psychiatric Diagnosis. In *Bias in Psychiatric Diagnosis*. P. Caplan and L. Cosgrove, eds. Lanham, MD: Jason Aronson.

Rawls, J. (1971), *A Theory of Justice*, Cambridge, MA: Harvard University Press.

Sanger, M. (1921), The Eugenic Value of Birth Control Propaganda, *Birth Control Review*.

Sederer, L. (2013), Involuntary Psychiatric Hospitalization: A Life Saving Yet Often Aversive Intervention, *Psychology Today*. URL= https://www.psychologytoday.com/blog/therapy-it-s-more-just-talk/201310/involuntary-psychiatric-hospitalization Retrieved January 1, 2018.

Sullivan, J. (2002), State Will Admit Sterilization Past, *Portland Oregonian*:A01.

Tekin, S. and Mosko, M. (2015), Hyponarrativity and Context-Specific Limitations of the DSM-5, *Public Affairs Quarterly*, 29:109–145.

Walker, J. (1993), *Women, Film, and Psychoanalytic Psychiatry*, Minneapolis: University of Minnesota Press.

20

The North African Syndrome: Traversing the Distance to the Cultural "Other"

Bryan Mukandi

> "Have I not, because of what I have done or failed to do, contributed to an impoverishment of human reality?"
>
> The question could also be formulated in this way:
>
> "Have I at all times demanded and brought out the man that is in me?"
>
> (Fanon 1967: 3)

I intend to dwell, in what follows, on movement and work; movement in the form of travel and labor. My purpose is the examination of the psychiatric interview where the clinician belongs to the dominant culture, whatever that means, and the client is a cultural "other"—an Indigenous person, an immigrant, a racial or ethnic minority. Who is traveling on what journey as these two are seated in the same room? What is the distribution of the labor demanded for understanding or intelligibility to arise between them? What might justice demand of them?

In their "Preface" to the third edition of *The Psychiatric Interview in Clinical Practice*, Roger MacKinnon, Robert Michels, and Peter Buckleymake the following illuminating claim: "The social relationship between patient and clinician *is no longer asymmetric*. Patients are now better informed, correctly believe that their bodies belong to them, and wish to be involved in treatment decisions. The therapeutic alliance between doctor and patient has become the foundation of treatment efforts in all of medicine" (MacKinnon et al. 2016: xi; emphasis added).

The claim is illuminating because of what I suspect is its prima facie validity, at least with respect to certain groups. There are groups of people for whom it may well be the case that what were once grossly asymmetric relations of power

with clinicians in psychiatric institutions have been dramatically leveled, aided in part by the greater power to choose between various types of psychiatric institutions, and manner in which one accesses their services.[1] On this view, abuses of power today are exceptions to the rule. In the normal course of affairs, not only is the "social relationship between patient and clinician" an egalitarian one, it is so for *everyone*: "*We now know* that the subjective experience of being 'different' is universal and that psychiatry is enriched by recognizing and exploring that experience, validating its existence and its universality, and attempting to understand how it influences the patient's life" (MacKinnon et al. 2016, emphasis added).

To repeat a friend's mantra, "Who is this we, that *we* are talking about?" Furthermore, what is it to *know* that "the subjective experience of being 'different' is universal"? Is it familiarity with a contemporary platitude? Or is it the instrumental knowledge of the sort one has when one knows lithium's pharmacological properties? If "*we*" know of the universality of "difference" in this instrumental manner, are we thereby necessarily armed with the ability to navigate and transcend that difference in the clinical encounter?

In a work of exceptional conceptual clarity, Julie Nordgaard, Louis Sass, and Josef Parnas suggest that in most psychiatric interviews, the "information provided by the patient coupled with his behavior, experience, and psychosocial history leads, in a natural conversational clinical situation … to the interviewer *seeing the patient as resembling a certain prototype*" (Nordgaard et al. 2013: 357; emphasis in original). This phenomenon, which these authors refer to as "the process of typification," is not limited to those clinicians who do not "know" about "difference."

> The process of typification refers to a very basic human cognitive feature, especially pertinent in perception, namely that perception of an object is always *apperceptively organized*, that is, structured, in a semi-conceptual fashion, as a salient unity or a certain Gestalt. In this sense, one might say, seeing is always "seeing as …"; it is always perspectival or aspectual. It involves pattern recognition and pattern completion, thereby allowing apprehension of objects and situations under conditions of limited or incomplete information. Typification is a largely automatic process that pervades all of our experiences and occurs outside explicit awareness … it imbues perception itself. (357; emphasis in original)

The attempt to recognize difference that MacKinnon and colleagues take for granted is subject to what Nordgaard and colleagues identify as "typification." That is, the clinical encounter is not a simple, transparent space. Nor is it discrete moment in time. It is a stage on which there are (at least) two actors, both engaged

in a long-running drama, whose performances demand interpretation. While this is fraught at the best of times, greater work is required of the clinician who would do no harm when the patient belongs to a different "cultural group." This is particularly the case when the clinician is a member of the "dominant group."

The encounters that I specifically have in mind are those between, for example, a white Euro-Australian psychiatrist and an Aboriginal and/or Torres Strait Islander client; or a white Swedish psychiatric nurse and a Turkish client. I want to explore the complexity of "typification" in cases such as these, and in so doing cast greater light on the vulnerability of clients who are culturally "other." My aim is to demonstrate just how much asymmetry there is, and how difficult it is to overcome, and to do that largely by retracing Frantz Fanon's "The 'North African Syndrome'" (1967), a text first published in 1952. While drawing on a text written in a different epoch in order to make claims about psychiatric practice in ours may seem anachronistic, my aim is to demonstrate the persistence of what can be understood as a "colonial ontologic architecture"—i.e., a framework circumscribing ways of being in relation to the other, where relational possibility is limited by the dictates of the colonial project.[2] In so doing, I aim to make clear the clinician's task when serving the "culturally other." I hope in the process to bring to light the significance of time, space, and movement in the intercultural psychiatric interview. To this end, I follow Fanon's lead, who structures his essays around three theses.

"First Thesis.—That the behavior of the North African often causes a medical staff to have misgivings as to the reality of his illness." (Fanon 1967: 4)

Fanon begins "The 'North African Syndrome'" by questioning the recognition of the humanity of "the North African who has emigrated to France" by members of French society, particularly medical personnel. "I ask you, I ask myself. Who are they, those creatures starving for humanity who stand buttressed against the impalpable frontiers (though I know them from experience to be terribly distinct) of complete recognition?" (Fanon 1967: 3). He illustrates the tragicomic manifestation of this lack of recognition in the clinical setting as follows:

> "What's wrong my friend?"
>
> "I'm dying, *monsieur le docteur*."
>
> His voice breaks imperceptibly.
>
> "Where have you pain?"
>
> "Everywhere, *monsieur le docteur*."
>
> …

> "Where does it hurt?"
>
> "In my belly." (He then points to his thorax and abdomen.)
>
> "When does it hurt?"
>
> "All the time."
>
> "Even at night?"
>
> "Especially at night."
>
> "It hurts more at night than in the daytime, does it?"
>
> "No, all the time."
>
> "But more at night than in the daytime?"
>
> "No, all the time."
>
> "And where does it hurt most?"
>
> "Here." (He then points to his thorax and abdomen.)
>
> And there you are. Meanwhile patients are waiting outside, and the worst of it is that you have the impression that time would not improve matters. *You therefore fall back on a diagnosis of probability* and in correlation propose an approximate therapy. (Fanon 1967: 4–5; emphasis mine)

There is no malice here on the part of the doctor—no hatred; no intention to do harm; not even clinical incompetence. There is, rather, the steadfast occupation of one's ground.[3] There is a manner in which *monsier le docteur* understands or *makes sense* of matters, and according to that manner of understanding, the North African client in question is not intelligible. That is, the burden of mutual intelligibility is not evenly distributed. *Monsier le docteur* makes sense of things in a particular manner, and the North African client in question does not register meaningfully on that basis. Faced with this lack of generation of intelligible meaning, the clinician makes "a diagnosis of probability." That is, the doctor finds a set of meanings that he already holds, and ascribes those to his encounter with his client. The phrase "to fall back upon" is misleading here since it connotes movement. However, what is striking is the doctor's *lack of movement*. He has the luxury to remain steadfast—to stand his ground. If the client's symptoms do not make sense on those grounds, it is to the client's detriment.

An analogous "case" is that of Nyasha, from Tsitsi Dangarembga's *Nervous Conditions*:

> In the city Maiguru's brother immediately made an appointment with a psychiatrist. We felt better—help was at hand. But the psychiatrist said that Nyasha could not be ill, that Africans did not suffer in the way we had described. She was making a scene. We should take her home and be firm with her. This

> was not a sensible thing to say in front of my uncle, who found these words vastly reassuring and considered going back to Umtali at once, turning a deaf ear to Nyasha when she begged to see an African psychiatrist …
>
> There were no black psychiatrists, but she was persuaded to see a white one. This man was human. (Dangarembga 2001: 207)

Nervous Conditions is set in the late 1970s, in what was then Rhodesia. That there were racist, white psychiatrists working in medical institutions at that time is not surprising. However, it is unwise to dismiss this incidence, or the one that Fanon chronicles, as either mere fiction, or worse, artifacts of a bygone era. When the psychiatrist suggests that Africans are unable to present with the signs and symptoms of certain conditions, what is on display is more than a set of preconceptions based on unfounded racist beliefs. More than that, what is on display is a particular manner of being in the world and apprehending or making sense of external phenomena. If one is convinced that unicorns do not exist but is then confronted with a unicorn, one's response betrays not one's attitudes or beliefs, but the process by which one *makes sense of* or *creates meaning* from experience. Nyasha's recourse to an African psychiatrist (which in this context means a black psychiatrist) has less to do with the assumption that an African is less likely to hold racist preconceptions, but more that such an individual will be more willing to do the work and undertake the movement necessary to understand Nyasha as a full human being and will make sense of her condition on those grounds.

Nyasha may not have found an African psychiatrist, but she does find another white one, one who is *human*. We will return to the significance of this. For now, let us focus instead on the fact that Fanon's client has no such good fortune. Fanon notes that:

> He does not go back to the same doctor, nor to the same dispensary. He goes elsewhere. He proceeds on the assumption that in order to get satisfaction he has to knock at every door, and he knocks. He knocks persistently. Gently. Naïvely. Furiously.
>
> He knocks. The door is opened. The door is always opened. And he tells about *his pain*. Which becomes increasingly his own. He talks about it volubly. He takes hold of it in space and puts it before the doctor's nose … It grows as one watches it. He gathers it over the whole surface of his body and after fifteen minutes of gestured explanations the interpreter (appropriately baffling) translates for us: he says he has a belly-ache. (Fanon 1967: 5, emphasis in original)

Fanon's client is aware of his body, and more specifically, the pain suffered by that body, as tragically his own—so his own that it is not communicable to the doctors that he consults, try as hard as he may. There is failure that occurs at the

site of translation. In crossing the threshold of intelligibility in the doctors' terms, that existential pain born of an inability to accommodate oneself to an unjust context is reduced to a "belly-ache," a complaint that one may expect of a child who has eaten a little too much or something that they ought not have eaten. The bafflement or perplexity induced by the interpreter is an allusion to chasm that separates the two parties—a chasm whose significance surpasses a language barrier. As Fanon goes on to explain, there is here "a kind of frustration in the field of explanation. The comedy or drama begins all over again: approximate diagnosis and therapy" (Fanon 1967: 6). How is this inability to make sense of this client's presentation, this recourse to "approximate diagnosis," made sense of by the clinician? The unreflected upon resolve to hold one's ground leads the clinician to "the negation of any morbidity. When you come down to it, the North African African is a simulator, a liar, a malingerer, a sluggard, a thief" (Fanon 1967: 7). That is, the clinician "falls back on" a set of meanings that they already have at their disposal. The problem of meaning is cast in terms of a problem that has to do with the "cultural other."

"Second Thesis.—That the attitude of medical personnel is very often an *a priori* attitude. The North African does not come with a substratum common to his race, but on a foundation built by the European. In other words, the North African, spontaneously, by the very fact of appearing on the scene, enters into a pre-existing framework." (Fanon 1967)

In considering "the attitude of medical personnel" toward their North African clients as described by Fanon, it is useful to recall that for Immanuel Kant, space and time are the *a priori* forms of intuition.[4] That is, "although the matter of all appearance is given to us only a posteriori, the form of all appearance must altogether lie ready for the sensations a priori in the mind" (Kant 1996: 73; A20|B34). What this means is that while we subsequently make sense of what is presented to our senses after the fact of that presentation, the mechanism by which it is possible to grasp sensory data is there before any specific encounter with an object. With respect to the concept of space for example, Kant writes:

> Space is a necessary a priori presentation that underlies all our intuitions. We can never have a presentation of there being no space, even though we are quite able to think of there being no objects encountered in it. Hence space must be regarded as the condition for the possibility of appearances, and not as a determination dependent on them. Space is an a priori presentation that necessarily underlies our appearances. (Kant 1996: 78; A24|B38–9)

Kant's understanding of time is similar, noting that "simultaneity or succession would not even enter our perception if the presentation of time did not underlie them a priori'" (Kant 1996: 85; A30|B46). What I want to complicate by citing these passages from Kant's "Transcendental Aesthetic" is the idea of immediacy. That is, the taken for granted attitude toward one's experience of objects in the world is one of direct contact. I see my book lying face down on the table and I pick it up, turn it over, and close the cover. It does not occur to me that there is an *a priori* framework along which that intuition, let alone cognition, takes place. Fanon's suggestion of an *a priori* attitude ought to call into question the taken for granted immediacy that one has in one's dialogue with another and in the encounter of doctor and client. His argument is that before *this particular individual*, who is a North African immigrant in 1952 Lyon, walks into the white, French psychiatrist's consultation office, there is a conceptual framework that is already at play, one that structures the perception and understanding of the medical practitioner. *This particular* human being who is seeking help is *a priori North African*. That is, the framework mediating the clinician's apprehension of the client is one that does not admit of that client as an individual, but as a North African where North African is a category imbued with an array of meanings.

Before following the implications of this situation in the clinical setting, an examination of the same phenomenon outside the clinical setting, as described by Fanon in a work published around the same period, will be helpful. Fanon described the fifth chapter of *Peau noir, masques blancs*, "*L'éxperience vécue du Noir*" as follows[5]:

> It shows *le nègre* (the nigger) face to face with their race … we witness the desperate efforts of *un nègre* (a nigger) who tries desperately to discover *le sens* (the sense or meaning) of black identity. White civilisation, European culture have imposed on the Black an existential deflection. We will show elsewhere that often what is called the black soul is a construction of the White. (Fanon 2011: 68; translation mine)[6]

This passage, tucked away in the book's "Introduction," is key, and we shall return to it. For now however, consider the manner in which this black individual comes "face to face with their race." Imagine an immigrant in 1952 Lyon sitting on a train who suddenly hears, "'Dirty nigger!' or simply 'Look! A Negro!'" (Fanon 2008: 89). This is how Fanon describes the impact of that encounter:

> I came into this world anxious to uncover the meaning of things ["*soucieux de faire lever un sens aux choses*" (Fanon 2011: 153)], my soul desirous to be at the origin of the world, and here I am an object among other objects.

> Locked in this suffocating reification, I appealed to the other so that his liberating gaze, gliding over my body suddenly smoothed of rough edges, would give me back the lightness of being I thought I had lost, and taking me out of the world put me back in the world. But just as I get to the other slope I stumble, and the Other fixes me with his gaze, his gestures and attitude, the same way you fix a preparation with a dye. (Fanon 2008: 89)

Fanon goes on to elaborate the process by which *this* black is subjected as *le nègre* in this racist social context. The protagonist is subjected in a myriad of ways to the notion that the "Negro is an animal, the Negro is bad, the Negro is wicked, the Negro is ugly; look, a Negro" (Fanon 2008: 93). It is inescapable. "The white man is all around me; up above the sky is tearing its navel; the earth crunches under my feet and sings white, white. All this whiteness burns me to a cinder" (Fanon 2008: 94). *This* black enters into a context in which there is already an established conception of *the* black.

> I slip into corners, my long antenna encountering the various axioms on the surface of things: the Negro's clothes smell of Negro; the Negro has white teeth; the Negro has big feet; the Negro has a broad chest. I slip into corners; I keep silent; all I want is to be anonymous, to be forgotten. Look, I'll agree to everything, on condition I go unnoticed. (Fanon 2008: 96)

Attempts to assert *this* black's particularity suffer the effects of a preexisting framework of understanding. Hence, while an individual may attempt to transcend this framework, even this attempted transcendence is understood in terms of that framework. The competent black physician is thereby deemed an exception to the rule, thereby maintaining the integrity of the framework. Yet even then, as Fanon's protagonist puts it, "I knew for instance if the physician made one false move, it was over for him and all who came after him. What, in fact, could one expect from a Negro physician" (Fanon 2008: 97). Fanon has the white world helpfully explain the situation to his protagonist: "You have come too late, much too late" (Fanon 2008: 101).

This predicament, understood by Fanon both spatially (being "fixed" as on a slide) and temporally (being too late) is described by Gordon as follows: "The schism between identity and being is destroyed, and the result is a necessary being, an overdetermined, 'ontological' reality. To see someone in this way is to close off possibilities. It takes the form of the command and the declaration instead of the interrogative" (Gordon 2015: 49).

If we return now to the passage above from the "Introduction," what is at stake is clear. *This* black person is made to come face-to-face with a conception of black identity in a context in which meaning-making belongs to the other.

That is, there are *a priori* ways of understanding or making sense to which *this* individual does not have access. *This* black cannot alter the manner in which black is understood and is therefore prone to being grasped in the same manner all blacks. "The Lived Experience of the Black" is a phenomenological account of this straightjacket. The "imposed deflection" to which Fanon refers is this being forced onto the rail tracks of an *a priori* system of understanding, where black identity or the black soul is determined along that route rather than in the terms that the protagonist might wish to give those things meaning. This ties in with Gordon's explanation if one considers Hans-Georg Gadamer's conception of understanding as having to do with one's ontological orientation. That is, for understanding to take place, one's situation must shift in order for there to be a "fusion of horizons"; a coming together of planes of understanding. "Transposing ourselves consists neither in the empathy of one individual for another nor in subordinating another person to our own standards; rather it always involves rising to a higher universality that overcomes not only our own particularity but also that of the other" (Gadamer 2004: 304).

As Charles Taylor notes however (and Fanon illustrates):

> The kind of understanding that ruling groups have of the ruled, that conquerors have of the conquered … has usually been based on a quiet confidence that the terms they need are already in their vocabulary … And indeed, the satisfaction of ruling, beyond the booty, the unequal exchange, the exploitation of labor, very much includes the reaffirmation of one's identity that comes from being able to live this fiction without meeting brutal refutation. Real understanding always has an identity cost—something the ruled have often painfully experienced. (Taylor 2002: 141)

Neither the interlocutor of Fanon's black, nor that of his North African client, move. Fanon's black is reduced to weeping by the end of the chapter. The North African is in pain. Of the black, Fanon writes: "At the start of my history that others have fabricated for me, the pedestal of cannibalism was given pride of place so that I wouldn't forget" (Fanon 2008: 100); of the North African:

> The North African today who goes to see a doctor bears the dead weight of all his compatriots. Of all those who had only symptoms … But the patient who is here, in front of me, this body which I am forced to assume is swept by a consciousness, this body which is no longer altogether a body or rather which is doubly a body since it is beside itself with terror—this body which asks me to listen to it without, however, paying too much heed to it—fills me with exasperation.

> "Where do you hurt?"
>
> "In my stomach." (He points to his liver.)
>
> I lose my patience. I tell him that the stomach is to the left …
>
> "It all hurts."
>
> I happen to know that this "it all" contains three organs; more exactly five or six. That each organ has *its* pathology. The pathology invented by the Arab does not interest us. It is pseudo-pathology. The Arab is a pseudo-invalid (Fanon 1967: 9).

The danger at this point is to imagine that what animates this scene is some regressive, false, fixed belief, however culturally acceptable it may have been in the time and place in which the scene is set. That would obscure the ways in which this scene is reproduced by progressive practitioners today. What establishes the conditions for the possibility of this exchange is that the clinician is able to steadfastly occupy the plane on which understanding takes place and meaning is ascribed (the clinician understands pathology and only what conforms to that understanding is legitimate). Consider what the client in question would have had to do in order to be understood. The client would not need to be familiar with human anatomy per se, but they would have had to come to an understanding of the clinician's understanding. That is, the client, the individual in pain or distress, must first undertake the movement necessary to be intelligible to the steadfast clinician. This is the work that the client who is "other" must do. The cross-cultural psychiatric interview is a field on which the Aboriginal or immigrant client is at work, carrying a disproportionate burden for the maintenance of intelligibility. If the Aboriginal or immigrant client fails to adequately carry this burden, they risk being taken for a pseudo-invalid, not only in Fanon's epoch, but also in ours.

Could it not be said that a "pseudo-invalid" is a "pseudo-person"? A full person demands of the clinician attention which the pseudo-person does not. Might one not therefore also suppose that in order to treat a client justly, a clinician must recognize that client's humanity—must grasp or appropriately fail to grasp the client as more than merely body "swept by a consciousness"; as human; as a person.[7]

"Third Thesis.—That the greatest willingness, the purest of intentions require enlightenment. Concerning the necessity of making a situational diagnosis." (Fanon 1967: 10)

Filippo Menozzi casts "the necessity of making a 'situational diagnosis'" in terms of the ability "to perceive the social condition of the North African immigrant worker: uprooted, often traumatized and marginalized" (Menozzi 2015: 366).

We again must ask what it means "to perceive" in this case. I do not think that Menozzi fully grasps the radicality of Fanon's demand, hence the former's reading of "the 'tension' or rupture between Fanon the psychiatrist and Fanon the activist" (Menozzi 2015: 368). I see no such rupture or tension. There is none if one takes the call for "a situational diagnosis" to be a call for the white, French psychiatrist to work; to move. It is a call to not only perceive in a removed, cerebral manner, but to undertake the ontological movement and perform the labor necessary to understand the client as a person, to understand them in their terms, according to the ontologic situation of the client—the ground that they occupy.

A consideration of the case of Salih F., from Miljenko Jergović's short story, "A Diagnosis," will illustrate what is at stake. "Salih F. saw with his own eyes his wife and two daughters being cut up with an electric saw by the Chetniks. Later, imprisoned in Manjača, he was expected to die there" (Jergović 1997: 108). He does not die in prison, but is released and eventually ends up in a refugee camp in the Czech Republic. Unfortunately, his time there is difficult:

> Illiterate and a bit slow, he was the ideal figure of fun. Salih F. spent days trying not to rise to the bait. Sometimes he made the effort to come up with a quick retort or to think of an original reply, but it didn't really work. He only ended up looking even more stupid than before. It was as though he had been dropped in a machine for mixing nerves. The only way to escape was to put the heat on the next sucker, or else to sort it out with his fists. (Jergović 1997)

Matters come to a head when Salih F. has "the shit beaten out of him, at first by the Bosnians and then by the Czech police guards" (Jergović 1997: 109). While the authorities want to deport him, they can neither send him back to a war-torn country, nor can they find a country willing to "take a Bosnian who [is] prone to fighting" (Jergović 1997). It is therefore finally decided to send him to a psychiatric hospital. Jergović describes the protagonist's situation there as follows:

> In hospital, Salih F. was treated like a king … The doctors were thrilled to have such an opportunity to study a human guinea pig who had witnessed his next of kin being cut to pieces—legs first, then arms, and finally the heads. At regular intervals the men in white coats used to check on Salih F. by peering through the spyhole … He looked no different from anybody else in the world who follows the latest news from Bosnia with a lack of interest.
>
> The doctors concluded that Salih F. was actually in a state of shock. They drafted long and pedantic reports about him, wrote papers for psychiatric journals, discussed various prognostications and cited patiently for Salih F.'s battered and bloody soul to recover. But over the next few months his condition did not change. (Jergović 1997: 109–110)

From the Fanonian perspective, Salih F.'s situation is similar to that of the North African and Dangarembga's Nyasha before the first psychiatrist that she consults. The humanity of all three goes unrecognized. Salih F. is reduced to a curiosity—an interesting case; a Bosnian where Bosnian has come to represent some cipher, some figure of lack or violence; something less than human. There are parallels between the attention that Salih F. receives and the inattention suffered by Fanon's North African client:

> "They can't express themselves."
>
> And they are liars
>
> and also they are thieves
>
> and also and also and also
>
> the Arab is a thief
>
> all Arabs are thieves
>
> It's a do-nothing race
>
> dirty
>
> disgusting
>
> Nothing you can do about them
>
> nothing you can get out of them
>
> sure, it's hard for them being the way they are
>
> being that way
>
> but anyway, you can't say it's our fault. (Fanon 1967: 14)

While the racism is blinding, Fanon's point is not that racist assumptions are wrong or irrational, or even that they are an impediment to good clinical practice. The point is that the starting point of a just clinical practice, the starting point of what Marilyn Nissim-Sabat calls "Fanonian humanism" (Nissim-Sabat 2010), is to engage in the work and to traverse the distance necessary in order to reckon with the other's humanity. Hence the charge that Fanon lays at his colleagues:

> This man whom you thingify by calling him systematically Mohammed, whom you reconstruct, or rather whom you dissolve, on the basis of an idea, an idea you know to be repulsive (you know perfectly well you rob him of something, that something for which not so long ago you were ready to give up everything, even your life) well, don't you have the impression of his substance? (Fanon 1967: 14)

Bearing in mind that this text was written less than ten years after the end of the Second World War, one can see that Fanon, the decorated veteran of that war, is assigning moral responsibility to his colleagues for their lack of recognition of the humanity of their North African clients. "As a person, you know very well what it is to suffer and to be in pain, and your lack of recognition of those states in your clients result from your decision to refrain from doing the work necessary to understand," he seems to say. This holds just as much for those attending to Salih F. and Nyasha. Even more radical is the responsibility that he places at the clinician's feet:

> There are houses to be built, schools to be opened, roads to be laid out, slums to be torn down, cities to be made to spring from the earth, men and women, children and children to be adorned with smiles.
>
> This means that there is work to be done over there, human work … there are tears to be wiped away, inhuman attitudes to be fought, condescending ways of speech to be ruled out, men to be humanized. (Fanon 1967: 15–16)

As Jean Khalfa illustrates in his excellent paper, Fanon recognized and took seriously the need to address the social and cultural factors that breed a variety of pathologies (Khalfa 2015). The work to which the clinician is tasked is not merely to travel the distance necessary to face the client as a person in their own right or on their own terms, but beyond that, to contribute toward the creation of a social and cultural context in which their humanity is recognized. His position is similar to Mogobe Ramose's understanding of the concept of *ubuntu*. Commenting on the Nguni proverb *umuntu ngumuntu ngabantu*, a person is a person by virtue of, among, and with other people, Ramose notes the following:

> Although the English language does not exhaust the meaning of this maxim or aphorism, it may nonetheless be construed to mean that to be a human be-ing is to affirm one's humanity by recognising the humanity of others and, on that basis, establish humane relations with them …
>
> … be-ing human is not enough. One is enjoined, yes, commanded as it were, to actually become a human being. What is decisive then is to prove oneself to be the embodiment of *ubu-ntu* … because the fundamental ethical, social and legal judgement of human worth and human conduct is based upon *ubu-ntu*. (Ramos 1999: 54–55)

The "enlightenment" or understanding necessary to make a "situational diagnosis" can only come about by engaging in something like Fanonian humanism—working toward the recognition of the humanity of the other, and in so doing, the expression of one's own humanity.

I have attempted to demonstrate the necessity of this disposition in the psychiatric interview, especially that in which the clinician belongs to the dominant culture and the client is what I have clumsily referred to as a "cultural other." In answer to the question of how realistic this is, or how one might to go about achieving this humanistic project, I leave you with Fanon's closing words on "The 'North African Syndrome'":

> Your solution, sir?
>
> Don't push me too far. Don't force me to tell you what you ought to know, sir. If YOU do not reclaim the man who is before you, how can I assume that you reclaim the man that is in you?
>
> If you do not want the man that is before you, how can I believe the man that is perhaps in you?
>
> If YOU do not demand the man, if YOU do not sacrifice the man that is in you so that the man who is on this earth shall be more than a body, more than a Mohammed, by what conjurer's trick will I have to acquire the certainty that you, too, are worthy of my love? (Fanon 1967: 16)

Notes

1 The lack of choice or power displayed in Miloš Forman's (1975) *One Flew over the Cuckoo's Nest* comes to mind. It is difficult to imagine the horrors viewed in that film being commonplace. This is compounded by Elaine Safer's (1977) demonstration, in an article titled "'It's the Truth Even if It Didn't Happen': Ken Kessey's *One Flew over the Cuckoo's Nest,*" of the manner in which the film mitigates the nightmarish quality of the book on which it is based.

2 Valentin Mudimbe describes the "colonial project" as one directed toward "the domination of physical space, the reformation of *natives*' minds, and the integration of local economic histories into the Western perspective" (1988: 2; emphasis in original).

3 While Fanon refers to Molière in this essay, there are also striking parallels between the psychiatrist-North African client drama that Fanon stages, and Samuel Beckett's tragicomedy *Waiting for Godot*, which was published in the same year as "The 'North African Syndrome.'" Consider the following, for example:

> *VLADIMIR* Well you see—
> *POZZO [peremptory]* Who is Godot?
> *ESTRAGON* Godot?
> *POZZO* You took me for Godot.
> *VLADIMIR* Oh no, Sir, not for an instant, Sir.

POZZO Who is he?
VLADIMIR Oh he's a.… he's a kind of acquaintance.
ESTRAGON Nothing of the kind, we hardly know him.
VLADIMIR True ….we don't know him very well ….but all the same ….
ESTRAGON Personally, I wouldn't even know him if I saw him.
POZZO You took me for him
ESTRAGON [recoiling before Pozzo] That's to say ….you understand ….the dusk ….the strain ….waiting ….I confess …. I imagined …. for a second ….
POZZO Waiting? So you were waiting for him?
VLADIMIR Well you see—
POZZO Here? On my land?
VLADIMIR We didn't intend any harm.
ESTRAGON We meant well.
POZZO The road is free to all.
VLADIMIR That's how we looked at it.
POZZO It's a disgrace. But there you are (Beckett 2010, pp. 71, 73; ellipses in original).

4 As Carrie Jenkins explains, *a priori* knowledge "is knowledge that is ….independent of any evidence gathered through sense experience ….knowledge whose justification does not consist in or ultimately rest upon empirical evidence" (Jenkins 2015, p. 207). That is, what is known *a priori* is known prior to experience, whereas the *a posteriori* is what is known following, or as the result of, experience.

5 In his translation, *Black Skin, White Masks*, Richard Philcox renders this chapter "The Lived Experience of the Black Man." I prefer "The Lived Experience of the Black" given the (nevertheless) problematic convention of ascribing to the masculine the status of universality. That is, grammatically, *le Noir* can refer to both the black man and the black woman in a way that *la Noire* (at that time) could not. I acknowledge however, that the question of the extent to which this chapter speaks to the lived experience of black women is a vexed, complex one. For further elaboration on this matter, I refer the reader to Gordon (2015) and Sharpley-Whiting (1998).

6 The original reads: "*Il montre le nègre en face de sa race ….nous assistons aux efforts désespérés d'un nègre qui s'acharne à découvrir le sens de l'identité noire. La civilisation blanche, la culture européenne ont imposé au Noir une déviation existentielle. Nous montrerons ailleurs que souvent ce qu'on appelle l'âme noire est une construction du Blanc.*"

7 For Emmanuel Levinas (1969), the *infinity* that is the other cannot be grasped or reduced to the *totality* that is my conceptualization of objects in the world. While they are very different in some ways, Fanonian and Levinasian clinical ethics are not incommensurable. Taken together, they demand that the clinician strive toward an understanding and grasp of the humanity of the client, without succumbing to the arrogance that one ever achieves this, since such "achievement" would amount to an Odyssian return to an *a priori* framework.

References

Beckett, S. (2010), *En attendant Godot/Waiting for Godot: Tragicomedy in 2 Acts; A Bilingual Edition Translated from the Original French Text by the Author*, New York: Grove Press.

Dangarembga, T. (2001), *Nervous Conditions*, London: Women's Press Classic.

Fanon, F. (1967), *Towards the African Revolution*: Political Essays. H. Chevalier, trans. New York: Grove Press.

Fanon, F. (2008), *Black Skin, White Masks*, R. Philcox, trans. New York: Grove Press.

Fanon, F. (2011), *Œuvres: Peau noir, masques blancs, L'An V de la révolution algérienne, Les Damnés de la terre, Pour la révolution africaine*, Paris: La Découverte.

Forman, M. (Director) (1975), *One Flew over the Cuckoo's Nest*, United States: Fantasy Films. Motion Picture.

Gadamer, H. G. (2004), *Truth and Method*, J. Weinsheimer and D. Marshall, trans. London: Continuum.

Gordon, L. (2015), *What Fanon Said: A Philosophical Introduction to His Life and Thought*, London: C. Hurst & Co.

Jenkins, C. (2015), A Priori Knowledge: The Conceptual Approach. In *The Bloomsbury Companion to Epistemology*. A. Cullison, ed. London: Bloomsbury.

Jergović, M. (1997), *Sarajevo Marlboro*, S. Tomašević, trans. London: Penguin Books.

Kant, I. (1996), *Critique of Pure Reason*, W. S. Pluhar, trans. Indianapolis, IN: Hackett Publishing Company.

Khalfa, Jean. (2015), Fanon and Psychiatry, *Nottingham French Studies*, 54:52–71.

Levinas, E. (1969), *Totality and Infinity: An Essay on Exteriority*, A. Lingis, trans. Pittsburgh, PA: Duquesne University Press.

MacKinnon, R., Michels, R., and Buckley, P. (2016), *The Psychiatric Interview in Clinical Practice: Third Edition*. Arlington, VA: American Psychiatric Association Publishing.

Menozzi, F. (2015), Fanon's Letter, *Interventions*, 17:360–377.

Mudimbe, V. (1988), *The Invention of Africa: Gnosis, Philosophy and the Order of Knowledge*, Oxford: James Currey.

Nissim-Sabat, M. (2010), Fanonian Musings: Decolonizing/Philosophy/Psychiatry. In *Fanon and the Decolonization of Philosophy*. E. Hoppe and T. Nicholis, eds. Lanham, MD: Lexington Books.

Nordgaard, J., Sass, L., and Parnas, J. (2013), The Psychiatric Interview: Validity, Structure, and Subjectivity, *European Archives of Psychiatry and Clinical Neuroscience*, 263:353–364.

Ramose, M. (1999), *African Philosophy through Ubuntu*, Harare: Mond Books.

Safer, E. (1977), "It's the Truth Even If It Didn't Happen": Ken Kessey's One Flew over the Cuckoo's Nest, *Literature/Film Quarterly*, 5:132–141.

Sharpley-Whiting, D. (1998), *Frantz Fanon: Conflicts and Feminisms*, Lanham, MD: Rowman & Littlefield Publishers.

Taylor, C. (2002), Gadamer on the Human Sciences. In *The Cambridge Companion to Gadamer*. R. J. Dostal, ed. Cambridge: Cambridge University Press.

Study Questions

1. *What is the relationship between power and psychiatry?*
2. *What are the implications of intersectional identity for the social status of people with mental disorders?*
3. *Is the diagnosis of mental disorder a form of oppression?*
4. *Is mental disorder real or is it a tool constructed for social control?*
5. *What are the implications of cultural differences on the individuation of mental disorders?*

Part Seven

Metaphysics, Epistemology, and Psychiatry

21

Mental Disorder, Free Will, and Personal Autonomy

Christian Perring

Introduction: The problems posed by mental illness for autonomy and freedom

A common philosophical project within the philosophy of psychiatry in investigating the nature of mental illnesses is to identify essential differences between people who are mentally ill and those who are not. In everyday thought, this is often done with regard to self-control. People with mental illness are often depicted as being out of control. This is sometimes expressed as being because the mental illness is controlling the person, so the person's actions are not freely chosen. This is especially true of addiction and obsessive-compulsive disorder (OCD), but is also true of schizophrenia, mania, and depression. This is one of the reasons that it is said that the actions of mentally ill people do not represent who they really are, or do not flow from their real selves. When it is claimed that mentally ill people are not responsible for some of their actions, it is the lack of self-control or freedom that is said to explain this.

This feature of our understanding of mental illness is so central that to deny it is to come close to denying that there are mental illnesses. If addicts and people with OCD, for example, turn out to be acting freely and with self-control, then we might find it hard to see why we should count these conditions as medical problems at all. We would be tempted to reclassify them as just reflecting a set of choices and preferences that we find hard to understand. So we would need to rethink how we understand mental illness if we cannot justify the idea that it indicates a lack of freedom and self-control.

However, it turns out that finding satisfactory theories of freedom or self-control that fit well with what we know about mental illness and explain our

commonly accepted assumptions about how the responsibility of mentally ill people is not easy. This has both *conceptual* and *empirical* sources. Providing a clear analysis of what we mean by freedom and self-control that fits with common intuitions has proven to be a challenge. What's more, the science of mental illness is also contested both within psychiatry and by psychiatric critics, and this complicates the task of matching up a conceptual analysis with empirical facts.

In this chapter, I will examine various approaches to understanding freedom and autonomy, and what kinds of mental illnesses entail a loss of freedom or autonomy. I will start with the metaphysical debate about free will, and move on to the debate about personal autonomy. I will finish with some reflections on how these debates could be settled, or at least, how one can come to a reasonable position in regard to understanding mental illness. I will especially focus on the methodological question as to whether we can let the social consequences of our views guide us as to which views to take.

Free will and determinism

In trying to conceptualize what mental illness takes away from us, we may be tempted to say that at least in severe cases, it takes away our freedom of action. Conditions of paralyzing anxiety or depression, or states of delusion interfere with people's lives to such an extent that we may suggest that they are less free as a result. We may explicitly say that at least with some illnesses such as addiction, people lose their freedom. We might then hope that philosophy could help to clarify what this could mean, and that we could use bring in philosophical theorizing to justify such claims.

Yet our intuitions about unfree action are elusive and maybe confused. We can distinguish between free action and forced action, for example, but we have to think hard about what counts as free action. We can deprive a person of their freedom to go where they want by locking them up in a prison cell, but we don't thereby force them to do anything. If a bully takes a victim's forearm and uses it to hit the victim's face, the victim isn't actively hitting themselves. If a mugger corners me in an alleyway brandishing a large knife, and tells me to hand over my wallet, we might say they force me to hand over the wallet, but we might equally say that I freely chose to hand over my wallet and I could have refused. When the local laws specify that I cannot park a car on the sidewalk, and there is a strong chance that I will be fined if I do so, it isn't obvious that they take away

my freedom to park on the sidewalk. They just give me an incentive to avoid such action. Similarly, the government provides me with a strong incentive to pay my taxes. The repercussions if I don't pay my taxes may be severe, and so we might say that we have to pay taxes or even that we are forced to pay taxes, but this may be an evocative way of making the point, while not being literally true. We could still refuse and accept the repercussions. So it proves difficult to find a crystal clear example of a person who is forced to perform an action. We can consider cases where the repercussions of not going along with what the person threatening us wants to do are increasingly unpleasant, involving our own pain, loss of property, or the infliction of pain on people we love, and maybe these asymptotically come to provoking the intuition that our action is forced.

We might also consider cases where our decision-making process is interfered with in other ways by other agents, through post-hypnotic suggestion, brain washing, brain implants, or magic spells. The central idea in this category is that the manipulator changes the victim's desires, so they behave differently, in accordance with the manipulator's plans. This might more plausibly be called unfree action, and we will discuss it further on. We can then discuss whether mental illnesses can be considered using this model, and so causing people to act unfreely.

This discussion brings out the one of the fundamental ideas about free action, that to be free, an agent could have done otherwise. Conversely, if an agent could not have done otherwise, then the agent was not acting freely. Thus if the agent's behavior is in some way determined by past events, then the agent's actions are not free. There is of course a vast literature in philosophy on free will and determinism. There are several camps of thought. A good way to divide them up is those who believe in determinism and those who don't, and this seems to be an empirical issue. For example, it seems that if some interpretations of quantum mechanics are right, then there are some events in the world that are genuinely random and are not determined by prior events. On the other hand, some claim that the world is deterministic (and maybe make a stronger claim that indeterminism is in some sense impossible). On this assumption, there are two main positions about free action. Incompatibilists say that determinism implies that people could not act other than they do, and that free action requires by its very nature that an agent could have done differently, so no actions are freely performed. Compatibilists respond that this is to misunderstand the meaning of free action, which simply requires that actions are not forced on a person. They say that determinism (of whatever

kind, be it causal, metaphysical, or of some other kind) is compatible with free action.

Some philosophers agree with the incompatibilist definition of freedom but argue the world is not completely determined, so argue that some actions are genuinely free in some greater sense than that of the compatibilist. One way of doing this is adopted by "libertarians" who say it is possible for agents to escape the chain of causation and act as an uncaused-causer, much as some theologians have argued that a supernatural being such as a Christian god can act freely.

None of this debate about determinism and indetermism sheds any particular light on the freedom or lack of it for mentally ill agents. Those who deny that there is any such thing as free will or free action obviously cannot draw a contrast between the freedom of people with mental illnesses and those without. For them, mental illness does not deprive people of the kind of freedom they are talking about because no one has that freedom.

Indeterminist libertarians might be in a position to mark some difference between those with mental illness and those without in regard to freedom, if they have an account of when a person can escape the causal chains of the universe. Needless to say, this would be a challenge. Certainly, no empirical work has indicated that some people escape the laws of causation, and so working out who has this special kind of freedom would be a matter of speculation. We would have no reason to think that people with mental illnesses have either more or less freedom than others on this account of freedom. It is possible that some accounts of some mental illness might appeal to ideas of randomness of causes of behavior, which might open the door to people with such conditions having more freedom than others. On this possibility, we might find an association between the erratic behavior of people with mental disorders and randomness of the causes, and this would be a result of their increased freedom. But it is hard to know how such an argument would actually proceed.

Generally, those who have held a compatibilist view of freedom and have argued that free action is that which is not compelled by others have not provided much clarification of what compelled action is, nor have they expanded on whether conditions such as addiction take away one's freedom. But as opposed to the view of the libertarians, there is a clear route to such discussion, because it is possible to explore in empirical and conceptual ways how action performed as a result of mental illness resembles (or is dissimilar from) action performed as a result of compulsion by others. It is this project that fits well with the philosophical debates over personal autonomy.

Personal autonomy

The central idea of free action is that one is not controlled by someone else or something external to oneself. This is equivalent to governing oneself rather than being governed by an external force. The task of a great deal of the literature on personal autonomy has been to set out in more precise terms what this could amount to, and what conditions undermine personal autonomy. One of the more recent debates is about what counts as an "external force." When the source of interference is another person, it is easy to understand how the interference is external to the person, but it has also been argued that it is possible to have mental states such as desires that are also external to oneself. Philosophical literature (stemming from Frankfurt's work on second-order desires and freedom of will, 1971) and psychological literature on ego-dystonic states (reference) have examined this. Some philosophers have argued that some actions do not stem from motives that the agent endorses, so the actions are in a sense not the agent's. A number of philosophers have attempted to elaborate how intentional actions can fail to belong to her. (Bratman 2007; Frankfurt 1971; Jaworska 2007; Shoemaker 2003; Watson 1975). The different views are similar in their approach but differ in details. Frankfurt argues that a person's true desires are the ones that they endorse reflectively, the ones they desire to desire, while the ones that they wish to be rid of do not fully belong to them. Watson argues in a similar vein that the true desires are those that the agent values. Bratman argues that for full attribution, an agent's desires must fit with their long-term plans. Both Jaworska and Shoemaker appeal to notions of what people care about as a criterion of which of their actions can be said to truly belong to them. All of these proposals overlap, which Buss (2018) labels as "coherentist," and while they may diverge in detail, they are motivated by very similar intuitions about people who find themselves with desires they do not endorse, and that autonomy requires being able to assess one's own motives for acting.

Other approaches to autonomy focus on whether the agent can respond appropriately to her reasons for action. Fischer and Ravizza (1998) argue that autonomy requires that an agent not just have coherence between her desires and values, but also that she be properly responsive to the world. She has to understand her situation in a way that her reasons are able to react to the world, so the person is not divorced from reality. While coherence is still important on this view, it is not sufficient for a person to be autonomous. An agent needs to have the appropriate connection to reality in order to be an effective agent, and

since autonomous agency is aimed at achieving one's goals, we need to include reasons responsiveness in the understanding of autonomy.

However, both these accounts, coherentist and reasons responsive, face the problem that while they may articulate ways in which desires and resulting actions can fail to be fully belong to a person, it is implausible that they specify necessary conditions for autonomy. While the analysis of weak-willed action is much debated, it is commonly accepted that people perform actions based on desires they do not endorse, and that they had self-control in doing do. Common examples of these are eating some unhealthy food after deciding to eat healthily, or snapping at someone after determining to be kind to them. To rule *a priori* that such actions are out of our control would take a good deal of philosophical argument and would go against our experience of them. But if this is so, then we have not achieved a strict criterion of autonomy.

On being forced by one's own desires

So we need to provide a further measure of when a person is unable to control their own actions that addresses the way that agents are unable to resist performing actions that they do not endorse. This has been attempted in both empirical and conceptual ways, with a broad overlap between the two. Unsurprisingly, philosophers have not achieved consensus on the best analysis of what counts as internally forced action, but more surprisingly, psychologists have also maintained a similar diversity of views and interpretations of empirical findings. This is probably a result of the difficulty of separating out empirical and conceptual components, and so any psychological claim about autonomy comes laden with philosophical assumptions. The debate in philosophy has focused mostly on the nature of substance addiction, but it has also included behavioral addictions, compulsions, tics, impulse control disorders, schizophrenia, depression, mania, and hyperactivity.

The debate over addiction gets muddied by the issue of whether addiction is a medical disease. Those who argue that it is tend to hold that addicts are not free in their addictive behavior, and those who argue that it is not a true disease tend to hold that active behavior is freely chosen. But there is no necessary conceptual tie between these joined positions, and the disease status of addiction is a distraction here.

A distinctive feature of the debate over addiction is how philosophical, psychological, neurological, sociological, and first person accounts co-exist,

overlap with each other, at the same time competing and complementing each other. It is often unclear when different views are being pitched at the same level of explanation and when they are offering different kinds of explanation. So when it comes to understanding the philosophical relevance of the many proffered theories of addiction, it is largely an open question how the empirical studies relate to the conceptual issues. (Levy's collection from 2013 amply demonstrates this. It is discussed at greater length in Perring 2011.)

Some argue that those with serious substance addiction lack self-control. This view is most often seen in popular discourse, and is generally stated in more qualified forms in scholarly literature. In the empirical literature, it is more often defended in neurological approaches, such as Volker and Fowler (2000). It is disputed as an empirical claim by many, notably Heyman (2009). Some approaches explain the irrationality of addictive behavior but do not address the issue of self-control per se. For example, in a series of papers, Ainslie (recently in his 2017) has used the resources of rational choice theory combined with the phenomenon of hyperbolic discounting to explain addiction as a kind of cognitive mistake resulting from habituation. Agents value the immediate satisfaction of preferences even when this is a costly choice. This violates rational choice theory, and is an unwise use of resources. This is a very different kind of explanation from saying that the person with the irrational behavior is overcome by the strength of their desires. While some authors write as if their proposed explanation of addiction is the only satisfactory explanation, it is likely that there are many different explanations, because addiction is a heterogeneous phenomenon with many different kinds of causal mechanisms—such as social, psychological, neurological, and genetic. There are probably many different phenomena grouped under the heading of addiction, brought together by a set of family resemblances rather than having a single essence in common.

In line with understanding the complexity of addiction phenomena and their explanations, it is also important to acknowledge that self-control and addiction are not all-or-nothing concepts. As Sinnott-Armstrong (2013) has recently argued, self-control comes in degrees: people can have more or less of it, and in different ways. The extent to which a person loses control of their own actions will vary according to the kind of causes of their addictive behavior. Assessing a person's responsibility for their actions will then need to be a more complex matter, with neither fully autonomy or full loss of autonomy being usual states of affairs.

I have emphasized the diversity of approaches to explaining addiction rather than spelling out any particular kind of explanation. We should note that some have attempted to bring some unity to the field with a synthesis of ideas. Holton and Berridge (2013) have attempted this, for example, setting out ordinary choice models, disease models, habit accounts, pleasure accounts, and learning accounts. They argue that they can be unified by distinguishing between the psychological states of wanting and liking, (where wanting is a motivational state that drives behavior and liking is a state of pleasure typically resulting from getting what one wants), and seeing addiction as a malfunction of the wanting system. Whether the field will move more toward consensus remains to be seen. We can hope that with a greater evidential base and more conceptual clarity, this would be likely, although looking at the rest of the field of psychiatry and the continuing proliferation of theories, we may have reasonable doubts that addiction is likely to buck the overall trend.

Cases of loss of autonomy

So far, we have seen some approaches to understanding free will and personal autonomy. I have emphasized the difficulties. Standard approaches to free will do not single out mental illness as a relevant category; theories of autonomy also do not provide clear and helpful accounts of self-control, and theories of self-control as applied to addiction have resulted in a proliferation of views with disagreements both on concepts and on interpretation of empirical evidence. We started this chapter with the observation for many mental illnesses, there is a strong intuition that they must involve a significant loss of freedom in order to count as illnesses or disorders at all. So the line of argument set out so far pushes us toward the conclusion that it is difficult to provide strong conceptual underpinnings to the view that most conditions classed as mental disorders are indeed disorders rather than patterns of unwise choices, problems in living, or some other state that makes life difficult for people.

This skeptical view is hard to square with the experiences of people with serious mental illnesses. People who are addicted to alcohol or drugs, for example, often testify that they tried many times to end their use of the substances but failed to do so, despite their best efforts.

Many other mental illnesses include symptoms which seem to involve actions which the sufferer would not otherwise choose. Here is a short list.

- In mania, people have great enthusiasm for projects that would not normally make sense to them, and they can engage in activities such as promiscuous sex or spending sprees that they would otherwise consider to be counter-productive.
- In compulsions that are part of OCD, people feel a need to perform actions in order to relieve an inner tension or to avoid some terrible consequence that they believe will follow if they don't perform the action. They may recognize in some way that the beliefs underlying the compulsion are baseless, but this generally does not reduce the need to act.
- With eating disorders such as binge eating, people eat large amounts of food which they know is bad for them, and experience great regret after the binge. It can then be followed by purging of the food.
- In some sexual disorders, people engage in sexual activity with which they are unhappy but which they still want because of the strong sexual excitement it brings them.

But note that not all these are experienced as going against the agent's values at the time. Indeed, sometimes the agent, at least at the time, wholly endorses their own actions. The action is not always experienced as alien or forced. This leaves us with the question of whether the behavior is autonomous or not.

From what we have seen already, it is already clear that for such confusing actions, it is very unlikely that the philosophical theories will deliver unambiguous answers. We need to lower our ambitions and investigate whether the theoretical apparatus can at least make some sense of the differences between the different kinds of action.

Let us consider some cases in a little more detail. One feature of mental disorders is that they last for some time, and at least according to the approach of the *DSM*, they cause problems in a person's life, interfering with work and relationships, even if the person with the mental illness does not necessarily accept that they have a disorder. When examining an action to see whether it was performed autonomously, we can distinguish three different classes of people with mental illnesses performing actions symptomatic of their illness.

1. Synchronic ambivalence. A concurrent judgment that one should not perform the action, held even as the person is engaged in the action.
2. Diachronic ambivalence. A judgment before and/or after the action that it would be a bad idea to do it, but no ambivalence while performing it.

3. Whole-hearted action. A person who is not ambivalent about the action when performing it, and who did not judge before or after that they should not do it.

We can stipulate that we are using the term *judgment* without strict definition here, acknowledging that there are many different kinds of judgment, and it can be a much broader category than just conscious belief formed in linguistic propositions.

On the coherentist views, it's fairly clear that synchronic ambivalence would entail that the desires leading to the action are disowned. This is not the same as being unfree action, but it gives us a way to say that the symptomatic desires do not belong to the agent in an important sense. With whole-hearted action, on this version of the coherentist view, the desires definitely belong to the agent. But different coherentists will probably diverge with the case of diachronic ambivalence. Frankfurt's view, for example, is largely about synchronic ambivalence, and so it would not count diachronic ambivalence as a threat to one's freedom. On the other hand, Bratman's view of intention is that it has to be understood within the context of one's longer term plans. Obviously, there are difficult questions of what length of time one needs to look at. Plans can change over long periods without a loss of autonomy. But if one has an aberrant desire that flies in the face of one's longer term plans, and one recognizes that one has these phases and aims to rid oneself of these aberrant desires, then this approach seems to provide one with a theoretical approach which allows us to say the desires of these phases are not completely attributable to the agent.

We get different results when we use the approach of Fischer and Ravizza. While their approach is complex, we can focus on the criterion of moderate reasons-responsiveness, which they hold is sufficient for moral responsibility. A person is responsible for their action when the mechanism that leads to their actions is moderately reasons-responsive. They spell this out as follows. Todd Long (2004: 152) provides a useful summary.

MRR: An agent's responsibility-relevant mechanism K is moderately reasons-responsive if and only if

(1) **K is regularly receptive to reasons, some of which are moral**; this requires (a) that holding fixed the operation of a K-type mechanism, the agent would recognize reasons in such a way as to give rise to an understandable pattern from the viewpoint of a third party who understands the agent's values and beliefs; and (b) that some of the reasons mentioned in (1.a) are moral reasons; and

(2) **K is at least weakly reactive to reasons**; this requires that the agent would react to at least one sufficient reason to do otherwise (in some possible scenario), although it does not follow that the agent could have responded differently to the actual reasons; and
(3) **K is the agent's own**; being the agent's own means "taking responsibility" for K; this requires that the agent (a) sees herself as the source of her behavior (which follows from the operation of K); and (b) believes that she is an apt candidate for the reactive attitudes as a result of how she exercises her agency in certain contexts; and (c) views herself as an agent with respect to (3.a)—(3.b) based on her evidence for these beliefs.

While this specification is quite detailed, we soon see the difficulties of applying to the cases of problematic agency above. It distinguishes between the agent and the agent's reasoning mechanism, which leaves open the possibility that an agent has more than one reasoning mechanism operating at different times or even concurrently, and there may be particular reason to invoke the idea of different mechanisms when a person is acting in bizarre ways due to mental illness. But let us just specify, in order for the theory to be usable at all, that a person has just one reasoning mechanism. For criterion (1) we need to judge whether the person is regularly receptive to reasons. Consider a person with OCD who washes their hands vigorously fifty times a day. People can often keep such symptoms well hidden, and other people will be unaware that the sufferer devotes hours each day to washing their hands. They can reason perfectly well about other parts of their lives, and they can often acknowledge the irrationality of their own handwashing. Regarding criterion (2) the agents are certainly reactive to some reasons, but are unmoved by the reasonable judgment that their hands are already clean enough and there are more important activities to be engaging in, such as spending time with their family. There may at least be some room here for interpreting the agent as being insufficiently reasons-responsive for them to be responsible for these actions. Finally, looking at criterion (3), this leads to a host of questions of what it means for a person see herself as the source of her behavior and an apt candidate for reactive attitudes such as blame. But again, there is probably room to argue that her experience is that it is the mental illness causing her action, and she does not see herself as a candidate for blame for her symptomatic behavior. While the details here need to be worked on more, we can see how this reasons-responsiveness view could be adapted provide a rationale for seeing symptomatic behavior as outside of the agent's responsibility.

The possibility of resolving the debate

We have examined a number of approaches to freedom and personal autonomy and their application to the actions of people with mental disorders. We have seen that general theories of freedom as related to determinism do not help to distinguish between people who are free and people who are not, but that different theories of personal autonomy do make such distinctions. So those theories are more helpful to us, and should be the kind of approach we are looking for.

However, we have also seen the difficulty in deciding which among them is the *right* theory. This should be no surprise: philosophy is not only difficult, but sociologically and historically, we know that philosophers tend to have long protracted disputes. *Epistemologically*, there is no certainty to be had with regard to which is the right answer. Nevertheless, we can still raise the issue of whether there is a right answer, even in principle—we can think of this as a *metaphysical* question. The protagonists in the debates certainly work on such an assumption, and aim to provide reasons for their answers. There is room for skepticism though. Here we can only briefly look at some considerations that include several unsubstantiated claims.

One method, or possibly a group of methods, to find the right theory of personal autonomy is through *conceptual analysis*. Exactly what constitutes this method is itself up for debate, but the basic idea is that we take our ordinary concepts of freedom and autonomy and carefully analyze them into a clear theory so that when we consider difficult cases such as actions by people with mental illness, we will find clear answers to when people are autonomous. However, given the profusion of different theories we have seen, there's reason to doubt that there is a pre-existing concept of autonomy that is detailed enough to provide answers to difficult cases. Furthermore, experience with discussing cases with a variety of people shows there is considerable variation in initial intuitions about cases. This casts further doubt on whether there is a shared concept of autonomy that will provide answers.

Alternatively, we might try to constrain the profusion of theories of autonomy by requiring them to be consistent with the *best scientific theories* we have. Indeed, we might hope that psychological and psychiatric science could give an indication of which theory of personal autonomy is the best fit. There is a good deal of work on agency and self-control in psychology, neuroscience, and psychiatry. Some philosophical work explicitly draws on this more empirical work. However, as explained earlier, there is a good deal of variation of views within the scientific theories of agency and self-control. It may be that the science

here is working toward a single unified notion of autonomy, but to assume that it must do so it to beg the question.

Such considerations point to the possibility that there will be continue to be many viable conceptions or theories of autonomy even as the empirical evidence accumulates and the philosophical debates proceed, and that they would disagree in major ways about when people with mental illnesses are autonomous. If this is right, then we cannot look to the philosophical debate over autonomy to dictate the right answers in real-world debates about how to understand mental illness. This is particularly important in the moral realm, where we might have hoped that the debate over autonomy would resolve questions of when people with mental illnesses are morally responsible for their actions.

So we should have modest expectations about what theories of autonomy can achieve in this area. As we have seen, they may provide insight into what kind of autonomy different people have, and they will show connections and contrasts between different phenomena of agency. Finally, we may decide to prefer a theory not because we have conclusive evidence it is right, but rather because for other reasons. For example, we might be wedded to a humanistic approach to people with mental illness which prioritizes their dignity and rejects dehumanizing or judgmental attitudes, and we may choose an approach to personal autonomy that is consistent with this approach. Thus, while many have argued that medical science has revealed that it is inappropriate to blame addicts for their behavior that causes many life problems for themselves and those close to them, the view we have arrived at here would advocate a similar conclusion for different reasons. Far from medical science forcing us to revelations about autonomy and responsibility that make blame inappropriate, it seems that it is possible to accept medical science and yet to hold a wide range of views about autonomy and responsibility. If we believe that it is morally preferable to avoid attitudes of blame, vengeance, and retribution toward people with addictions, this could push us to choosing models of autonomy that best fit with this generous nonjudgmental moral approach. The moral choice would thus drive our metaphysical view of autonomy and addiction, rather than be dictated by the metaphysics.

References

Ainslie, G. (2017), Palpating the Elephant: Current Theories of Addiction in Light of Hyperbolic Delay Discounting, pp. 227–244. In *Addiction and Choice: Rethinking the Relationship*. N. Heathe and G. Segal, eds. New York: Oxford University Press.

Bratman, M. (2007), *Structures of Agency: Essays*, New York: Oxford University Press.

Buss, S. (2018), Personal Autonomy. In *The Stanford Encyclopedia of Philosophy*, Winter 2016 edition. E. N. Zalta, ed. URL= https://plato.stanford.edu/archives/win2016/entries/personal-autonomy/

Fischer, J. and Ravizza, M. (1998), *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge: Cambridge University Press.

Frankfurt, H. G. (1971), Freedom of the Will and the Concept of a Person, *The Journal of Philosophy*, 68:5–20.

Heyman, G. (2009), *Addiction: A Disorder of Choice*, Cambridge, MA: Harvard University Press.

Holton, R. and Berridge, K. (2013), Addiction between Compulsion and Choice. In *Addiction and Self-Control: Perspectives from Philosophy, Psychology, and Neuroscience*. N. Levy, ed. New York: Oxford University Press.

Jaworska, A. (2007), Caring and Internality, *Philosophy and Phenomenological Research*, 74:529–568.

Long, T. R. (2004), Moderate Reasons-Responsiveness, Moral Responsibility, and Manipulation, pp. 151–172. In *Freedom and Determinism*. J. Keim-Campbell, M. O'Rourke, and D. Shier, eds. Cambridge, MA: MIT Press.

Perring, C. D. (2011), Bridging the Gap between Philosophers of Mind and Brain Researchers: The Example of Addiction, *Mens Sana Monograph*, 9:193–201.

Shoemaker, D. (2003), Caring, Identification, and Agency, *Ethics*, 114:88–118.

Sinnott-Armstrong, W. (2013), Are Addicts Responsible? In *Addiction and Self-Control: Perspectives from Philosophy, Psychology, and Neuroscience*. N. Levy, ed. New York: Oxford University Press.

Volkow, N. D. and Fowler, J. S. (2000), Addiction: A Disease of Compulsion and Drive: Involvement of the Orbitofrontal Cortex, *Cerebral Cortex*, 10:318–325.

Watson, G. (1975), Free Agency, *Journal of Philosophy*, 72:205–220.

22

Metaphysical Problems in Psychiatric Classification and Nosology

Peter Zachar

Introduction

Metaphysics is one of the most speculative and controversial topics in philosophy. Traditionally, it has been understood as the study of the nature of reality and of what exists. For example, the question of whether the universe is completely material or whether the universe is mostly empty space containing material particles was an important metaphysical dispute up until the early twentieth century. According to one current scientific metaphysics, most of space approximates a vacuum (i.e., is devoid of matter) and matter is ultimately composed of one-dimensional extended objects called strings.

Another example of a metaphysical claim is: *All that is real is reasonable, and all that is reasonable is real*. In the middle years of the twentieth century, a group of thinkers called the logical positivists believed that such claims played no useful role in understanding the world—declaring that all metaphysical claims are meaningless (Carnap 1936; 1937; 1967). By this they meant that there was no observable state of affairs in the world that would make these claims true. In their view, any meaningful claim about reality would either be a logical truth (e.g., a = a) or be one that was verifiable by observation, and thereby scientific.

The philosopher of science Karl Popper (1963) disagreed with the logical positivists. According to him, metaphysics is not science, but nor is it meaningless. For example, the claim *all events have a cause* is considered metaphysical, but also meaningful. Many scientific research programs depend on such metaphysical claims.

The logical positivists' dismissal of metaphysics was also undermined after their views about the nature of observation in science were shown to be

inaccurate (Suppe 1977). For instance, astronomers have learned that stars are mostly composed of hydrogen and helium, but stars also contain small amounts of heavier elements, such as carbon and oxygen. The scientific observation of a star's composition, using a technique called spectroscopy, depends on complex theories about electrons, electron shells, and the absorption and emission of photons. Such technological approaches to observation have called attention to the importance of theories in making observations, including metaphysical theories.

Popper's colleague, Imre Lakatos (1970), extended this rehabilitation of metaphysics by suggesting that scientific paradigms (or research programs) are always guided by metaphysical assumptions. Lakatos said that all research programs have a "hard core," by which he meant a set of assumptions that are metaphysical because they cannot be questioned without abandoning the research program itself. Research programs also have auxiliary metaphysical assumptions that can be tested and modified, which Lakatos called the positive heuristic. For example, in the seventeenth century, the metaphysical theory that space is a material plenum and vacuums are impossible was tested and disproved by Robert Boyle (Shapin and Schaffer 1985).

Many philosophers have come to believe that implicit metaphysical commitments are unavoidable heuristics that play a guiding role in the sciences. One of the tasks of philosophy, therefore, is to identify implicit metaphysical assumptions in scientific research programs. A second task, if possible, is to improve metaphysical theories—usually by making them more consistent with each other and with the evidence.

This chapter examines important metaphysical issues in the philosophy of psychiatry with respect to classification and nosology. There are too many different metaphysical problems inherent to psychiatric classification and nosology to survey here, so I have selected three topics to explore in more detail. These are ontology, causality, and the nature of clinical entities. These topics were selected because they are directly relevant to ongoing debates about classification in psychiatry and clinical psychology.

Ontology

The fundamental question of metaphysics—what is there or what exists—is called ontology. In psychiatry and clinical psychology the fundamental metaphysical question is whether mental disorders exist.

For much of recorded history people have harbored a suspicion that those with mental disorders are not actually ill. They might be considered criminal, lazy, or weak-willed instead of ill. For nearly as long, however, those who advocated for a moral rather than a medical viewpoint have been considered either uninformed or lacking in compassion.

A much more serious challenge to the reality of mental illness was offered by the psychiatrist Thomas Szasz (1961). Szasz was not lacking in compassion, but he believed that the concept of mental illness was a metaphor used to describe certain problems-in-living. He accepted that these were real problems, but not real medical illnesses.

According to Szasz, a valid medical illness is a disease defined by a physical lesion or alteration of some biological norm. No such lesions have been discovered for mental illnesses, and if they were ever discovered, Szasz believes that these disorders would be classified as brain diseases and treated by neurologists. Any problem that is "mental," such as thinking about killing oneself in the midst of a depressive disorder, violates cultural and social norms, not biological norms. Mental *illness*, claimed Szasz, is a myth.

Szasz's assertion caused quite a stir—for many reasons, not the least of which was the implication that psychiatrists are agents of social control. For example, through much of the twentieth century, Russia and all its surrounding states were organized into a governmental body called the Soviet Union. The Soviet Union did not allow a free press or have protections for free speech, and citizens were not allowed to publically disagree with the government. Some political dissidents who publically criticized the government were diagnosed as suffering from "sluggish schizophrenia" and confined to an institution. According to the Soviet psychiatrists, only someone mentally ill could oppose Soviet policies (van Voren 2010).

Something similar occurred in the United States. Before slavery was outlawed in 1865, a southern physician named Samuel Cartwright ([1851]2004) argued that black slaves who made repeated attempts to escape to the free states did so because they had a mental disorder called drapetomania. He claimed that drapetomania made black slaves unable to accept their natural state of servitude. Cartwright also believed that this "disease" could be cured. With proper management by their white masters, black slaves would never develop the pathological desire to run away.

In these cases, the diagnosis of sluggish schizophrenia and drapetomania were tools of social control that enforced the values of the local culture. According to the Szaszian view, all psychiatric diagnoses are tools of social control.

This raises some interesting questions. Is something a mental illness only because psychiatrists decide to treat it based on their social norms and values? If so, we could not argue that the Soviet dissidents and runaway slaves were not really disordered because disorders are defined relative to local norms. Nor could one condemn a community of psychiatrists who declared that women who demanded the right to vote to or become scientists were mentally disordered if women were seen as naturally unsuited to voting and higher education in those psychiatrists' culture.

Using such as relativist definition, nearly any violation of a social norm could be labeled as a mental disorder. Thus was born the important philosophical problem of defining the nature of mental disorder in an objective, nonrelativist way.

The naturalist perspective

The naturalist perspective disputes the claim that mental illness is constituted by violations of social norms and values (i.e., normativism). Naturalists agree that people with a mental disorder violate social norms, but that is not what makes them ill. What makes them ill is the presence of an objective disease process.

In a book titled *The Reality of Mental Illness*, Martin Roth and Jerome Kroll (1986) pointed out that diseases such as epilepsy, diabetes, and syphilis were recognized as diseases long before any biological lesions were discovered. The same could be said for most dementias and autoimmune diseases. Physicians did not need to see biological lesions to understand that these were diseases, nor do psychiatrists to see lesions to understand that depression and schizophrenia are mental disorders.

According to Roth and Kroll, the Szaszian claim that real diseases consist of observable lumps, lesions, or infections is based on a narrow and faulty concept of disease. In fact, some treatable "diseases" such as high blood pressure and elevated levels of cholesterol are merely elevated levels of normal physiological measurements that represent risks to health, not pathological entities.

Roth and Kroll also argue that the dramatic examples of psychiatrists being agents of social control, such as Soviet Psychiatry, were never considered legitimate by the psychiatric profession and were widely condemned as a misuse of psychiatry. Indeed, there is extensive continuity across time and across cultures about psychosis, depression, and debilitating anxiety. Rather than being singled out as targets of social control, these problem-in-living came to be seen

as medical because both the person with the problem and their family members could no longer manage on their own and sought help.

Roth and Kroll propose than rather than lesions and lumps, what unites the variety of conditions called diseases is that they all interfere with functioning and thereby place people at a biological disadvantage.

The notion that diseases are conditions that place people at a biological disadvantage was originally articulated by John Guyett Scadding (1967) and expanded on by Robert Kendell (1975). After medical doctors accepted that normal variations such as high blood pressure could be diseases, the physical lesion model of disease was replaced by a statistical model in which diseases are united by being deviations from what is normal.

The problem is that many deviations from normal, such as high levels of intelligence, are not diseases. The biological disadvantage view says that only those deviations from normal that lead to reduced fertility and higher levels of mortality (i.e., biological disadvantages) are diseases. For instance, consistent with the biological disadvantage model of disease, those with schizophrenia are less likely than most people to have children; they also die at a younger age than normal (Kendell 1975).

In the philosophy of medicine the leading advocate for the naturalist approach to disease is Christopher Boorse (1975, 1977). Boorse's view is based on the concept of *natural function*. According to him, natural biological functions are those functions that contribute to survival or reproduction. Natural psychological functions do the same. Disease refers to any internal state that interferes with a natural function.

In the Boorsian view, the norms violated by mental disorders are not social and cultural norms as Szasz claimed, but statistical norms. What is normal is the statistically typical contribution that a natural function makes to survival or reproduction. According to Boorse, if the heart was designed in evolution to pump blood, then a typical heart will pump blood. If the memory system was designed in evolution to store and recall important episodes in people's lives, then the typical memory systems will store information and allow for its recall. What is natural, therefore, is also typical or statistically normal.

Boorse states that although we tend to think of diseases as bad and unwanted things, being evaluated negatively is not what makes anything a disease. What counts as statistically normal functioning is purely a factual matter. Whether some internal state interferes with that function's contribution to survival or reproduction is also a factual matter. For Boorse, whether something is a real disease is matter of objective fact, and not a value judgment.

The fact + value hybrid model

Jerome Wakefield (1992a, 1992b), a trained mental health professional and a philosopher, introduced a hybrid model of mental disorder in which disorders must have an objective, fact-based component, but they must also violate social and cultural norms.

Wakefield agrees with Boorse that natural functions and dysfunctions are matters of fact, and not based on value judgments. According to Wakefield natural functions were selected during evolution because they increased our ancestors' chances of surviving. Their contribution to survival explains why they exist. Dysfunctions, he says, are failures of naturally selected functions.

Wakefield's notion of natural function, however, is more historical than Boorse's notion. For instance, according to Wakefield, our liking for both fat and sugar increased our ancestor's chances of survival in environments where reliable access to high-quality nutrition was infrequent and therefore such taste preferences represent natural functions. However, in an environment with ready access to food, including manufactured foods such as bread, pasta, and chocolate, our natural taste preferences increase our chances of being overweight and developing diabetes and heart disease. A liking for fat and sugar is maladaptive in our current environment but not a dysfunction because it is in accord with our biological design.

Not only can natural functions be maladaptive, dysfunctions can be adaptive. For instance, we are designed to have two hands, with five fingers on each hand. Being born with seven fingers would represent a dysfunction. If a fourteen-fingered person were a pianist, however, she could probably play complex pieces that no ten-fingered musician could replicate. Having fourteen fingers would be a dysfunction with an adaptive advantage.

According to Wakefield, the presence of a dysfunction is necessary for something to be a disorder, but it is not sufficient. Dysfunctions must also be harmful to their bearer in some way, with harmful meaning an impairment of personal, occupational, or social functioning; that is, there must be a violation of cultural norms and values. It is because our pianist would not be harmed by having fourteen fingers that we would not view her as having a disorder.

In Wakefield's model, major depressive disorder must be a dysfunction in the person, but it must also be harmful to its bearer. As noted, the kinds of harms that matter in psychiatry are disturbances of personal functioning (such as loathing one's self), occupational functioning (such as missing work), and social functioning (such as not taking care of one's children).

An important goal of Wakefield's *harmful dysfunction model* is to call attention to mis-diagnosing normal conditions as mental disorders. For example, even if disagreeing with the government or seeking to escape slavery are disvalued by a particular society, neither of them is a failure of a naturally selected function and therefore not a mental disorder.

One last point deserves mention. The concept of disease refers to many different things, including infectious diseases, genetic diseases, and acquired heart and lung diseases. Diseases also form only a part of treatable medical conditions, which also include injuries, pains, deformities, disabilities, and threats to health. These are often lumped together as "disorders." Rather than mental illness, psychiatrists and psychologists more typically talk about "mental disorder," "psychiatric disorder," or "psychopathology."

Causality

Earlier in this chapter I stated that causality is considered a metaphysical concept. This may surprise some people—what can be more obvious than the fact of causality? Upon reflection, however, the reason that causality is a topic in metaphysics is readily understandable.

Consider what would happen if you held a glass aloft and then let go of it. It would fall to the ground. But why would it fall? Most people will answer that gravity caused the glass to fall to the ground. However, the philosopher asks—have you ever *seen* gravity? If so, what color is it? Does gravity have a shape? Can gravity move? The obvious answers to these questions are that gravity has no color, shape or state of motion. We never "see" gravity. Nor do we see cause.

Writing about causality David Hume ([1739]2000) analyzed what we do see. For one, we see the glass held aloft and we see it on the ground. So we see *before and after*. Also, the letting go of the glass and the falling occur in the same location and at the same time. So we see *spatial and temporal contiguity* between cause and effect. In addition, every time I hold the glass a loft and let it go, it falls to the ground. So we see regularity. In his discussion of regularity, Hume says that letting go of the glass and its falling to the ground are *constantly conjoined*. Before and after, spatial and temporal contiguity, and constant conjunction, however, do not constitute a metaphysical cause that makes something happen.

Hume famously added that our idea of causality also includes the notion that the glass did not fall to the ground by accident—it was not a random occurrence. As a matter of fact, if someone was holding the glass aloft and it slipped out of

their hand, they would immediately know that it was going to fall to the ground before they saw it fall, and would be surprised if they let go of the glass and it floated away. Another word for nonrandom, or something that did not occur by chance, is "necessary."

According to Hume we cannot observe necessity or the "making happen" part of cause. Based on repeated experiences of the same pattern, we make an automatic inference about the relationship between cause and effect which we experience as a feeling (or impression) of causal necessity. By feeling or impression Hume means a subjective sense of a connection—the immediate knowing that the glass will fall referred to earlier. An example of how causes and effects are automatically tied together in our minds is our tendency to use "letting go of the glass" and "letting the glass drop" synonymously, but in fact letting go is a before and dropping is an after.

For Hume, our metaphysical notion of causes as making things happen is constructed out of something from the world and from the mind. From the world there is the observation of regularity, from the mind there is an impression of necessity related to an automatic, habitual inference.

Laws versus causes

Hume intended his work on causality to be consistent with Isaac Newton's ideas about gravity. For Newton, the law of gravity was a statement about a regularity that obtains between two material objects: they have a mutual attraction that is directly proportional to the masses of the objects and inversely proportional to the distance between them. Newton did not attempt to explain why gravity existed or how it worked. For him it was just a regularity. Following Newton and Hume, many scientists sought to discover such regularities—with the most important regularities being called laws of nature. Examples of hypothesized laws in the science of psychology include *the law of effect* and the *Yerkes-Dodson law*.

An important feature of laws of nature is that they are universal and admit no exceptions. The law of gravity works the same everywhere in the universe. However, no biological species on Earth—including human beings—is universal in the way gravity is. Biological species do not exist everywhere and most of them eventually go extinct. For example, consider the human heart. There are no laws of nature that necessitate how a heart must be structured or even that living organisms must have hearts (i.e., single-cell organisms do not have hearts). In biology, variation and exceptions are as much the rule as is sameness. One can

formulate many useful generalizations about species, but these generalizations do not rise to the level of laws of nature. If we want to explain the structure of the human heart we have know about its evolutionary history and the selection pressures that guided its development.

Turning to those features of the human species studied by psychology, let us look at the so-called Yerkes-Dodson law. According to this law, performance improves as level of arousal rises, to a point, after which performance declines. For instance, worrying about a test is beneficial because it motivates you to study and thereby perform better, but if your worry becomes too intense, your performance will decline.

Unlike gravity, this "law" does not apply to rocks, plants, or amoebas. Even where it does apply, there is great variation. For tasks at which we are not skilled and lack confidence, even a small amount of arousal can lead to worse performance and this will vary for the same person over time. For some people, on tasks on which they are highly skilled, they may never experience arousal great enough to effect performance. The Yerkes-Dodson principle is an informative empirical generalization, but not quite a universal law.

The upshot is that in contemporary psychology there is little interest left in discovering laws. The new ideal goal is causal explanation. For instance, we might want to explain why the Yerkes-Dodson generalization works and why it breaks down.

As a medical discipline, the search for causes has been more important in psychiatry than psychology, however, causal explanations have not been easy to confirm as valid, in part, because in psychiatry causality is complicated. For example, consider the claim that child abuse is a causal factor in the development of borderline personality disorder. If so, child abuse is a philosophically strange kind of a cause because the abuse presumably occurred years before the disorder developed. In Hume's notion of cause, the cause precedes the effect, but there is also a close temporal relationship in which the effect occurs soon after cause appears, not years later. The glass fell to the ground immediately after it was let go, not several years later.

A similar problem occurs in thinking about genes as causes—some of whose effects often manifest long before psychiatric disorders appear. Genes are causes of psychiatric disorder more like tobacco is a cause of cancer. Not everyone who smokes gets cancer, and not every with the relevant genetic predispositions develop psychiatric disorders.

In addition, some common causal claims about psychiatric disorder may not even refer to actual causes. Consider the claim that depression is caused by

a chemical imbalance in the brain such as a low level of the neurotransmitter serotonin. A very old question in philosophy is whether a neural state, such as a chemical imbalance, can cause a psychological state, such as depression. One common answer is that depression is a psychological description of a particular kind of brain state. If so, then the brain state is not the cause of the depression any more than the wood is the cause of the table or hydrogen and oxygen are the causes of water. The table is wood and water is H_20. If the chemical imbalance hypothesis is correct, then a low level of serotonin is depression, not its cause.

Although examples such as glasses falling to the ground help us understand what we mean by the concept of cause, they fail to represent all the types of causes recognized in science, including and especially the causes of psychiatric disorder.

Causal pluralism

It is generally accepted that psychiatric disorders are the result of a plurality of causes, rather than single cause (Kendler 2012). For instance, genetic predispositions, temperament, personality traits, coping styles, hormones and blood sugar levels, recent stress, levels of social support, cultural and economic conditions, and much more are all part of what determines someone's mood. Studying any one of these factors could help us understand a person's ongoing mood, but none of them is the cause of mood—instead the whole causal field is what determines mood.

One way of dealing with the complexity of causal fields was proposed by John Mackie (1965) in the form of the INUS condition. Philosophers traditionally demarcate causes in different ways. For instance, Causes can be *necessary* for an effect, for example, being infected with the TB bacillus is necessary for having tuberculosis. They can also be *sufficient* for an effect, for example, beheading someone is sufficient for killing them, but not necessary because you can kill them in other ways as well. How do these criteria apply to effects resulting from a plurality of causes? According to Mackie, what we practically call "the cause" is often an Insufficient but Necessary part of an Unnecessary but Sufficient condition, i.e., an INUS condition.

This a mouthful, but the idea is that there are many different factors that influence whether an outcome occurs. Consider depression. In a person with a past history of depression, being prone to negative emotions and lacking social support in combination with the death of a loved one can initiate a recurrent depressive episode. Rather than calling this whole field "the cause,"

we can practically identify the loss of a loved one as the cause because it is an INUS condition. On its own, the loss of a loved one is not sufficient to cause depression, but without it, the other factors in the causal field would not have resulted in this person becoming depressed. In Mackie's terms, causes are things that are necessary for a particular combination of factors to produce an effect, but they do not have to be necessary or sufficient for the effect by themselves.

Mackie's account is Humean in a primary sense because causes are identified with observable patterns or regularities. It is also Humean in a secondary sense because according to Hume, practically speaking, causes are difference makers—which means that if the cause had not occurred, then the effect could not have occurred. The INUS condition is a difference maker.

Another practical approach to causality in the regularity tradition, proposed, James Woodward (2008), is called the interventionist perspective. This perspective views causes as things we can manipulate and thereby alter effects. According to Woodward, X is the cause of Y if we can intervene on X and thereby change Y. For instance, in the case of a grief-related depression, it may be true that if the person did not lose a loved one, they would not be depressed, but after such a loss occurs we cannot do anything to bring back the person who died. Let us say we can intervene on negative emotionality or intervene on social support and improve the person's relationships. If either of these interventions leads to a reduction in depressive symptoms, then these factors are playing a causal role. If the interventions fail, then we are not intervening on causes.

This, however, is complex. We may find that once a depressive episode becomes severe, the person's mood is not responsive to any change in the environment, so altering the level of social support would not make them less depressed. Social support is not occupying a causal role in that circumstance. For another person, however, if their mood is responsive to the environment and they are depressed following an interpersonal rejection, an increase in social support might make them less depressed. So in this circumstance social support would be playing a causal role.

In contrast to social support, if intervening on negative emotionality leads to a reduction in depression for most people, then negative emotionality is a particularly useful cause because its causal efficacy is invariant across different circumstances. Such interventions are not invariant in the way that laws are, but they are more or less invariant. According to Woodward, interventions that are invariant across changing conditions are most useful.

What is also interesting about Woodward's interventionist theory of cause is that psychological variables can also count as causes. Let us say we can use

psychotherapy to alter depressogenic beliefs such as *It is unfair when things do not go like I want them too; I must always succeed or I am worthless*; and *I fail at everything I try.* If intervening on these beliefs improves the mood of people who are depressed, then the beliefs count as causes. Any X is a cause of Y if you can intervene on X and change Y.

Let us say that we can also use antidepressant medication to intervene on a pathological brain state to improve mood in depressed people. Should we say that the medical intervention is better because it is targeting the real cause of the depression?

Not necessarily. For one, the assertion that brain states are the *real* causes is metaphysically loaded. Brain states clearly play important causal roles, but adding the adjective "real," represents a highly speculative metaphysical augmentation. According to the interventionist model, if an intervention alters the effect, then the target of that intervention is a real cause.

Furthermore, which causal factor is better is a practical question. The problem with antidepressant medication is that it may also cause weight gain, loss of interest in sex, and blurred vision, that is, it causes too much for our purposes. If intervening on beliefs has lasting effects on depression, and does not require tolerating unpleasant medication side effects nor being on medication for years, then practically speaking, beliefs are better causes than brain states.

The nature of clinical entities

Sameness

What does it mean to say that depression runs in a family and a patient has inherited the same depression that one or more of his parents had? When we say that a person experiences a second or third episode of depression, is this the recurrence of the same depression from before? What is basis of this sameness?

One answer to this question is called the theory of universals. Consider the property of being a triangle. What makes different triangles the same this is that they exemplify the universal property of triangularity. The property of triangularity is equally present in everything that is triangular. Triangularity is also a part of the ontological structure of the world because even if we could get rid of all actual triangles, that would not destroy triangularity.

If depression is a universal, then different cases of depression exemplify the same thing—sometimes called the essence or nature of depression. Furthermore if psychiatrists were able to prevent depression from ever occurring, we would

continue to have knowledge of the universal so it would not be destroyed. If later in human history people became depressed again, it would be the same depression that was previously eliminated.

According to the theory of universals, every actual instance of depression may differ slightly, just like all actual triangles differ. To understand the universal, however, is to understand depression and triangularity in their pure forms.

The contrast to the theory of universals is called nominalism. According to one popular version of nominalism, depression is a general name for the way that people with depression resemble one another, but is not literally some spooky property that can be fully present in more than one place at once. One can say, loosely, that this person inherited his father's depression or his depression has recurred, but these depressions are different things. They are the same in the way that two people might wear the same dress to a wedding, not the same in the way that two people can live in the same house.

Properties and entities

A universal is usually considered a property, for instance the ball is red, Celeste is intelligent, and John is depressed. In this case, redness, intelligence, and depression are properties exemplified by the ball, Celeste, and John, respectively. But major depressive disorder also has properties such as feeling sad and weight loss, which are also universals. As properties, reported symptoms (such as sadness) and observable signs (such as weight loss) are fallible indicators of an underlying pathological entity.

With respect to the metaphysics of entities, properties are ways the entity appears, but it is the entity that grounds the properties. Although the symptom presentation for major depressive disorder may vary slightly from case to case, they all are grounded by the same pathological entity.

Entities that ground properties are called substrates. According to philosophers such as John Locke, however, we think that substrates ground properties, but our notion of a substrate may be nothing more than a speculative "we know not what" that grounds properties. Locke accepted that substrates exist, but considered them unknowable. According to Hume, however, because the properties are all we ever know, substrates are better treated as metaphysical fictions. What exists are clusters of properties.

For instance, the properties of psychiatric disorders are symptoms. Is it possible to be clinically depressed or psychotic but have no symptoms? That does not mean having symptoms of depression but acting as if all was normal, it means having no absolutely symptoms that are associated with suffering or

functional impairment. No sadness, hopelessness, weight loss, psychomotor retardation, and so on. It may be possible that depression results from an underlying dysfunction that does not always manifest as symptoms, but if a person cannot have an asymptomatic psychiatric disorder, then from a Humean perspective, the important clinical entities are not underlying dysfunctions (i.e., fictional substrates), but clusters of correlated signs and symptoms.

In metaphysics, advocates for universals generally prefer the substrate approach to entities. If properties such as being red or dysphoria are universals that can occur in more than one place, then they have to be grounded in something particular that occurs in only one place, and that is the substrate.

Most nominalists are skeptical about spooky universals and hypothetical substrates. Many of them prefer a property cluster approach to entities. Every property and every property cluster is a particular thing occurring in one place and has a beginning and an ending. For example, one person's depression is a different particular than another person's depression. They may resemble each other enough to form a class, but they are distinct existences.

Psychiatric entities: three metaphysical views

The property cluster view raises the question of why the symptoms of depression co-occur, which takes us right back to something hidden behind the appearances that makes the co-occurring symptoms a coherent entity. Three distinct, yet metaphysically committed views on clinical entities are represented by the medical model, the factor analytic model, and the causal network model.

According to the medical model, signs and symptoms co-occur because they are manifestations of an underlying pathological process. The pathological process is the real clinical entity. This pathological process may be either biological or psychological in nature.

A related but different approach to the co-occurrence problem has been developed in clinical psychology. Psychologists agree that correlations between the symptoms occur because of a hidden causal factor, but they call this hidden factor a latent variable. Using the statistical approach called factor analysis, psychologists can determine if the correlations between symptoms can be explained by latent variables.

There are many ways to statistically model the psychiatric symptoms space, but factor analysis has tended to outperform other approaches (Helzer et al. 2008). Factor analytic models, however, do not carve the symptom space into diagnostic categories such as depression and schizophrenia, they carve it into

latent variables. The latent variables derived in factor analysis are *dimensions*. Dimensions are continuous variables on which everyone has a value from low to high. Examples of dimensions are height and weight.

Some of the most metaphysically committed advocates of the dimensional model are proponents of the five-factor model of personality and personality disorder (McCrae and Costa 1997). The five personality factors are extroversion, neuroticism, agreeableness, openness, and conscientiousness. Its leading advocates declare that these five dimensions represent the universal structure of human personality:

> why are there five factors … ? We believe it is simply an empirical fact, like the fact that there are seven continents on earth or eight American presidents from Virginia. Biologists recognize eight classes of vertebrates (mammals, birds, reptiles, amphibians, and four classes of fishes, one extinct), and the theory of evolution helps to explain the development of these classes. It does not, however, explain why eight classes evolved, rather than four or eleven, and no one considers this a defect in the theory. There are, of course, reasons why human beings differ along each of the five personality dimensions—reasons to be found somewhere in evolution, neurobiology, socialization, or the existential human condition. But it is probably not meaningful or profitable to ask why there happen to be just five such dimensions. (McCrae and John 1992: 194)

An influential factor analytic model of major depressive disorder is called the tripartite model (Clark and Watson 1991). It is composed of three independent dimensions named positive affect, negative affect, and physiological hyperarousal. "Independent" means that a person's position on one of these three dimensions is uncorrelated with her position on the other dimensions. These dimensions also span across the normal and the abnormal. For instance, every person always occupies a position on the continuous dimension of positive affect—ranging from absent to very high. Both very low positive affect (anhedonia) and very high positive affect (mania) are psychiatrically abnormal. According to the tripartite model, major depressive disorder is a name we give to the combination of low levels of positive affect, high levels of negative affect, and low levels of arousal.

The advocates for dimensional models believe that diagnostic categories such as depression and schizophrenia are based on clinical tradition and opinion, while factor analytic dimension such as positive affect and negative affect are scientifically derived. As we saw earlier in the comparison between continents and personality factors, some thinkers believe that the dimensions discovered by factor analysis are the real "entities" of psychiatry (Livesley 2012; McCrae

and John 1992). According to this view, major depressive disorder is not a real clinical entity, but a name we give to a profile of high and low values across a set of dimensions.

A third approach to the co-occurrence problem is called the causal network model. According to this model, symptoms of depression such as sadness, sleep disturbance, fatigue, and concentration problems are correlated, but rather than reflecting the influence of underlying or hidden variables, the correlations represent direct causal connections between the symptoms themselves. For instance, not sleeping can cause one to feel fatigued and in turn to have concentration problems (Cramer et al. 2010).

From this causal network perspective, to develop a depressive episode is to enter a particular symptom network. For instance, there is some evidence that insomnia and depressed mood have many causal connections with other depression symptoms and when they occur in a vulnerable person, they initiate the formation of the interlocking network called a depressive episode (Cramer et al. 2010).

Often, a depressive episode may be triggered by a stressful event or an important loss. This reactive depressive episode may represent a temporary disruption that soon dissipates. If the depressive symptom network becomes strongly connected, however, it will be sustained by feedback loops even after the triggering events subsides, and the reactive depressive episode will be converted into major depressive disorder (Borsboom 2017).

According to the causal network approach, the real clinical entity is the symptom network. Without the correlations between the symptoms, there would be no clinical entity. Rather than being fallible indicators of an underlying pathological processes or a latent dimensions, the symptoms are parts of the casual network (McNally et al. 2015).

What do the network model advocates have to say about the metaphysical claims of the medical model and the dimensional model? With respect to the medical model, they argue that psychiatrist have been unable to find defining underlying disease mechanisms for mental disorders because such mechanisms do not exist (Borsboom 2017). With respect to the dimensional model, they argue that the dimensions/latent variables of factor analysis are scientific fictions. Indeed the advocates for network models have shown that the statistical technique of factor analysis will churn out latent variables like positive affectivity even if, in reality, the correlations between the variables represent not hidden variables, but causal relationships between the manifest variables themselves (Borsboom and Cramer 2013; van der Maas et al. 2006).

To conclude, we have just examined three metaphysical views on the nature of clinical entities. According to the medical model, a mental disorder is an underlying pathological process that manifests through symptoms. According to the factor analytic model, a mental disorder is a configuration of high and low scores on a collection of dimensions that are relevant to psychiatric distress and impairment. According to the network model, a mental disorder represents the stable state of a strongly connected symptom network.

Conclusion

This chapter has offered only a small sampling of the metaphysical issues that are important to psychiatric classification and nosology, and only a brief treatment of this small sample. Let me close by noting that the topics studied in metaphysics are rarely independent of one another, so once you commit to a position on one issue; it has implications for your views on other issues. Sometimes a strong metaphysical commitment on one issue will constrain what you can consistently accept on another.

If, as suggested at the beginning of the chapter, research programs always contain metaphysical assumptions, and, as argued at the end, those assumptions have implications for other metaphysical assumptions, then the role of metaphysics in science—including psychiatry—may be more extensive than is often considered. But so considered, metaphysics refers to the most general assumptions made in science, and should be seen as potentially continuous with science, not separate from or necessarily incompatible with it.

References

Boorse, C. (1975), On the Distinction between Disease and Illness, *Philosophy and Public Affairs*, 5:49–68.

Boorse, C. (1977), Health as a Theoretical Concept, *Philosophy of Science*, 44:542–573.

Borsboom, D. (2017), A Network Theory of Mental Disorders, *World Psychiatry*, 16:5–13.

Borsboom, D. and Cramer, A. O. J. (2013), Network Analysis: An Integrative Approach to the Structure of Psychopathology, *Annual Review of Clinical Psychology*, 9:91–121.

Carnap, R. (1936), Testability and Meaning, *Philosophy of Science*, 3:419–471.

Carnap, R. (1937), Testability and Meaning-Continued, *Philosophy of Science*, 4:1–40.

Carnap, R. (1967), *The Logical Structure of the World and Pseudoproblems in Philosophy*, R. A. George, trans. London: Routledge & Kegan Paul.

Cartwright, S. A. ([1851] 2004), Diseases and Physical Peculiarities of the Negro Race, pp. 28–39. In *Health, Disease, and Illness*. A. L. Caplan, J. J. McCartney, and D. A. Sisti, eds. Washington, DC: Georgetown University Press.

Clark, L. A. and Watson, D. (1991), Tripartite Model of Anxiety and Depression: Psychometric Evidence and Taxonomic Implications, *Journal of Abnormal Psychology*, 100:316–336.

Cramer, A. O. J., Waldrop, L. J., Van Der Mass, H. L. J., and Borsboom, D. (2010), Comorbidity: A Network Perspective, *Behavioral and Brain Sciences*, 33:137–150.

Helzer, J. E., Kraemer, H. C., Krueger, R. F., Wittchen, H. U., Sirovatka, P. J., and Regier, D. A. (2008), *Dimensional Approaches in Diagnostic Classification: Refining the Research Agenda for DSM-V*, Arlington, VA: American Psychiatric Association.

Hume, D. ([1739] 2000), *A Treatise of Human Nature*, Oxford: Oxford University Press.

Kendell, R. E. (1975), The Concept of Disease and Its Implications for Psychiatry, *British Journal of Psychiatry*, 127:305–315.

Kendler, K. S. (2012), The Dappled Nature of Causes of Psychiatric Illness: Replacing the Organic-Functional/Hardware-Software Dichotomy with Empirically Based Pluralism, *Molecular Psychiatry*, 17:377–388.

Lakatos, I. (1970), Falsification and the Methodology of Scientific Research Programmes, pp. 91–196. In *Criticism and the Growth of Knowledge*. I. Lakatos and A. Musgrave, eds. New York: Cambridge University Press.

Livesley, W. J. (2012), Tradition versus Empiricism in the Current DSM-5 Proposal for Revising the Classification of Personality Disorders, *Criminal Behavior and Mental Health*, 22:81–91.

Mackie, J. L. (1965), Causes and Conditions, *American Philosophical Quarterly*, 2:245–264.

McCrae, R. R. and Costa Jr., P. T. (1997), Personality Trait Structure as a Human Universal, *American Psychologist*, 52:509–516.

McCrae, R. R. and John, O. P. (1992), An Introduction to the Five Factor Model and Its Implications, *Journal of Personality*, 60:175–215.

McNally, R. J., Robinaugh, D. J., Wu, G. W., Wang, L., Deserno, M. K., and Borsboom, D. (2015), Mental Disorders as Causal Systems: A Network Approach to Posttraumatic Stress Disorder, *Clinical Psychological Science*, 36:836–849.

Popper, K. (1963), *Conjectures and Refutations: The Growth of Scientific Knowledge*, London: Routledge.

Roth, M. and Kroll, J. (1986), *The Reality of Mental Illness*, Cambridge: Cambridge University Press.

Scadding, J. G. (1967), Diagnosis: The Clinician and the Computer, *Lancet*, 290: 877–882.

Shapin, S., Schaffer, S. (1985), *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*, Princeton, NJ: Princeton University Press.

Suppe, F. (1977), Introduction, pp. 3–232. In *The Structure of Scientific Theories*, 2nd edition. F. Suppe, ed. Urbana: University of Illinois Press.

Szasz, T. S. (1961), *The Myth of Mental Illness*, New York: Harper & Row.

Van Der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., and Raijmakers, M. E. J. (2006), A Dynamical Model of General Intelligence: The Positive Manifold of Intelligence by Mutualism, *Psychological Review*, 113:842–861.

Van Voren, R. (2010), Political Abuse of Psychiatry—An Historical Overview, *Schizophrenia Bulletin*, 36:33–35.

Wakefield, J. C. (1992a), The Concept of Mental Disorder: On the Boundary between Biological Facts and Social Values, *American Psychologist*, 47:373–388.

Wakefield, J. C. (1992b), Disorder as Harmful Dysfunction: A Conceptual Critique of DSM-III-R's Definition of Mental Disorder, *Psychological Review*, 99:232–247.

Woodward, J. (2008), Cause and Explanation in Psychiatry: An Interventionist Perspective, pp. 136–184. In *Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology*. K. S. Kendler and J. Parnas, eds. Baltimore, MD: Johns Hopkins University Press.

23

Scientific Realism, Antirealism, and Psychiatric Diagnosis

Sam Fellowes

Introduction

Taking the *Diagnostic and Statistical Manual of Mental Disorders (DSM)* and the *International Classification of Diseases (ICD)* as archetypes, a psychiatric diagnosis primarily lists symptoms and signs of people whom it purports to describe. It may also list additional factors such as how the condition develops over time, prevalence figures, causal factors, and gender differences. In this paper I focus on the behavioral aspects of psychiatric diagnosis. There is a long history of concern over psychiatric diagnosis. Some critics consider psychiatric diagnosis as arbitrary or as a means of pathologizing people with only minor problems. They sometimes argue that, in contrast to typical pictures of science as providing descriptions of how the world actually is, psychiatric diagnoses have been invented, created, or constructed by psychiatrists. This is important because psychiatric diagnoses can influence decisions over treatment and over compulsory detentions. Being diagnosed can influence an individual's self-perception and their relationship to others. Psychiatric diagnoses being somehow false or untrue would raise serious questions over how much they should influence clinical decision making. I will provide an overview of arguments for and against psychiatric diagnosis meriting belief through exploring notions of realism and antirealism present in analytical philosophy of science and philosophy of psychiatry.

Questions of realism are about whether our beliefs accurately describe the world. To the degree that our beliefs accurately describe the world, realism is justified. To the degree that our beliefs do not accurately describe the world, then, a form of antirealism is justified. For example, most people are realists

about tables. When most people speak about or use tables it seems there is something which exists. In contrast, most people are antirealists about unicorns. These are typically understood as being imaginary. They are something we have created or constructed in our minds.

This chapter focuses on a particular type of realism, scientific realism.[1] This topic focuses on the questions about the realism or antirealism of scientific concepts. The central issue of the scientific realism debate is whether scientific theories correctly or very nearly correctly describe the world. Most people generally believe that what scientists describe actually exists. Some philosophers, known as scientific realists, share these views. Other philosophers are scientific antirealists. They commonly point out that there have been many scientific theories which scientists believed in, sometimes on very good evidence, yet those theories were mistaken in many of their central claims. Despite making incredibly successful predictions many past abandoned theories postulated things which turned out to actually not exist. A scientific antirealist, Larry Laudan, lists many theories which scientists once believed in, but which turned out to be mistaken in central details about what exists in the world:

-the crystalline spheres of ancient and medieval astronomy;

-the humoral theory of medicine;

-the effluvial theory of static electricity;

-"catastrophist" geology, with its commitment to a universal (Noachian) deluge;

-the phlogiston theory of chemistry;

-the caloric theory of heat;

-the vibratory theory of heat;

-the vital force theories of physiology;

-the electromagnetic aether;

-the optical aether. (Laudan 1981: 33)

The realist and antirealist debate thus aims to establish whether modern scientific theories are similarly mistaken in central claims about what things exist in the world, questioning if things like electrons and tectonic plates actually exist or if they are mistaken in a similar manner to old, abandoned scientific theories. The history of psychiatry also includes many long-abandoned psychiatric diagnoses such as hysteria, monomania, childhood schizophrenia, neurasthenia, draeptomania, simple schizophrenia, and symbiotic psychosis, raising such antirealist concerns toward psychiatric diagnosis.[2]

Scientific realist Stathis Psillos outlines scientific realism as having two aspects, both of which must be met for a scientific theory to merit belief. Failure to meet either of them results in antirealism, whereby a scientific theory does not merit belief. These aspects are metaphysical realism and epistemic realism. Metaphysical realism focuses on questions about what exists. Epistemic realism focuses on questions about how much we know about what exists. The remainder of this chapter first considers arguments for and against metaphysical realism in psychiatry. That is, what part, if any, of the real world could psychiatric diagnosis describe? Then the chapter secondly considers arguments for and against epistemic realism in psychiatry. That is, how reliable are psychiatric diagnosis for describing what exists in the parts of the world psychiatry investigates?[3]

Metaphysical realism and antirealism

Scientific realism involves "*The Metaphysical Thesis*: The world has a definite and mind-independent structure" (Psillos 2009: 4). A metaphysical realist thinks the world has a mind-independent structure which will be in principle knowable. The world has a particular form independent of our views about it. It exists in a particular way, and there are facts about the way the world exists, regardless of how we think the world is. Metaphysical realists see the mind-independent structure of the world as what is real, as what scientific theories try to literally describe. In contrast, metaphysical antirealism is where particular facts about the nature of the world are only present after human conventions are in place. A metaphysical antirealist denies there is a mind-independent structure. Astrology is a fairly uncontroversial example of something toward which philosophers would adopt metaphysical antirealism. Astrology makes claims that things like the positions of the planets can influence the interpersonal relationships of humans. If the actual metaphysical structure of the world is one in which the position of the planets does not affect human interpersonal relationships, then astrology does not study the metaphysical structure of the world. Astrology attempts to describe something which has no metaphysical structure so the relationships posited by astrology describe nothing real.

Metaphysical realism and antirealism are often domain specific. Someone might think there are true facts about the nature of the world in parts of the world studied by some sciences but not in areas studied by other sciences. Some examples highlight how philosophers could be metaphysical realist about some scientific claims but be a metaphysical antirealist over other claims. A

philosopher might believe there are mind-independent facts about atoms in physics and cells in biology but not about organisms. For a second example, a philosopher might deny that there are mind-independent facts about economic activities of nations. There might be facts about people, geographic locations, and exchange of money but not about nations. A metaphysical antirealist might claim nations are arbitrary groups, in that it is purely human conventions which decide where the boundaries between Germany and France are. There are no mind-independent facts about economic activities of nations because which nations there are is purely a matter of arbitrary human conventions. Thus a metaphysical antirealist about psychiatric diagnosis does not believe that nothing exists. Rather, they might be metaphysical realists about many different scientific claims but not over psychiatric diagnosis.

The central question here is whether there is any part of the metaphysical structure of the world which psychiatric diagnosis might describe. The alternative is that, like with astrology, psychiatric diagnoses are things which humans have created and that do not describe the real world. I shall now outline metaphysical realism over psychiatric diagnosis then I shall discuss arguments for metaphysical antirealism over psychiatric diagnosis.

A metaphysical realist over psychiatric diagnosis will think that the symptoms (thoughts, feelings, and behavior) exhibited by patients is part of the mind-independent structure. Alternatively, the metaphysical realist over psychiatric diagnosis will think that those symptoms (thoughts, feelings, and behavior) is caused by the mind-independent structure of the world. A metaphysical antirealist will deny these claims. The most commonly employed notion of metaphysical realism in psychiatry is that of "natural kinds." These are typically thought of as natural divisions in the world which form due to real causes. Since "membership in a natural kind depends on the common basis that holds the kind together, we have a ready solution to the problem of why patients who meet the diagnosis share a set of symptoms, viz. that they are all undergoing the same processes and the symptoms are effects of those processes" (Murphy 2014: 106). On this account, individuals with the same underlying physiological characteristics will exhibit the same behavior. The same causes in the metaphysical structure of the world are present in each individual with the diagnosis. Therefore, a psychiatric diagnosis would merit belief if it correctly described that behavior. There are, however, many different approaches to natural kinds. I will first outline four different accounts of kinds in psychiatry, as presented by Nick Haslam. Then, I will assess in what manner each one is or is not realist.

First, Haslam outlines essentialist natural kinds where a "specific, causally efficacious pathological process, mechanism or structure is present" (Haslam 2002: 212). This is essentialist because the underlying cause is present for everyone with that illness and absent in those without. Also, essentialist natural kinds have clear boundaries. Someone either has that psychiatric diagnosis or they do not; no one partially has the diagnosis. Secondly, Haslam outlines discrete kinds, which also have clear boundaries, meaning someone cannot partly meet the diagnostic criteria. However, discrete kinds differ from essentialist natural kinds because discrete kinds do not have essentialist causes (Haslam 2002: 210). On this account, multiple different causes can come together to produce a discrete set of symptoms; there is no need for one single cause to be present. Thirdly, Haslam outlines fuzzy kinds which lack the discrete boundaries of discrete kinds. There is no sharp natural demarcation between where the fuzzy kind starts and ends. While there is a difference between some individuals with fuzzy kinds and some individuals who lack that fuzzy kind, there are also cases where individuals have only some of the symptoms of the psychiatric diagnosis so partly have the fuzzy kind and partly do not (Haslam 2002: 208). Additionally, they also lack essentialist causes. Finally, Haslam outlines practical kinds whereby there is no boundary in the world between those with the psychiatric diagnosis and those without. Rather, the boundary is "imposed for practical reasons [rather than being] detected because it is in some respect objectively present" (Haslam 2002: 214). Whereas people with a particular fuzzy kind are separated from those who do not have it (even if there are people who only partly have it), with practical kinds there is no clear separation between anyone with a practical kind and anyone without it. The place where a particular practical kind starts and ends is decided by psychiatrists rather than determined by the world.

Following Haslam's four different types of kinds, psychiatric diagnoses can clearly come in many forms. Which of these forms count as realist? Philosophers disagree about this matter. I consider two different realist accounts of natural kinds by philosophers of science and match them to Haslam's four different types of kinds. Brian Ellis developed a notion of natural kinds named scientific essentialism (Ellis 2001: 22). According to Ellis, there are naturally forming entities in the world which have essential properties. All instances of a particular entity have the same properties so will be identical to each other. Additionally, since the essential properties of one entity are different to the essential properties of another entity those two entities form a natural demarcation in the world. This is roughly Haslam's notion of essentialist natural kinds. Philosophically informed psychiatrists Robert Kendell and Assen Jablensky took quite a similar position,

arguing that valid psychiatric diagnoses have necessary and sufficient biological mechanisms, ones always present in that illness and not present in other illness. Any psychiatric diagnosis not meeting these criteria will be arbitrary (Kendell and Jablensky 2003: 9). Down's syndrome is one of the few psychiatric diagnoses to fit essentialist natural kinds, each individual with Down's syndrome having a specific underlying cause and everyone without Down's syndrome lacking that specific underlying cause.

A more popular realist account of natural kinds in philosophy of psychiatry employs Richard Boyd's homeostatic cluster kind approach. In this account there are causal mechanisms in nature which then produce clustering sets of characteristics. "The natural definition of one of these homeo-static property cluster kinds is determined by the members of a cluster of often co-occurring properties and by the ('homeo-static') mechanisms that bring about their co-occurrence" (Boyd 1991: 141). This position differs from Ellis's scientific essentialism on two grounds. First, there is no requirement that the homeostatic mechanisms are essentialist. Not every mechanism will be present in all cases of the disorder. Secondly, the characteristics produced by these mechanisms need not form into discrete bundles. Some homeostatic property cluster kinds will be sharply demarcated from other homeostatic property cluster kinds, but not all will be. Boyd's account fits Haslam's account of discrete kinds and fuzzy kinds. Homeostatic mechanisms can produce groupings of symptoms which are closely correlated, the symptoms being either present or absent in an individual, and thus produce discrete kinds. However, sometimes homeostatic mechanisms produce groupings of symptoms which occur together in a looser, more irregular fashion, the symptoms only being partially present in some individuals, and thus produce fuzzy kinds. Many philosophers of psychiatry employ Boyd's account of natural kinds (Beebee and Sabbarton-Leary 2010; Kendler et al. 2011; Murphy 2014; Tsou 2016).

I now bring together Haslam's four types of kinds with these two realist notions of kinds. Haslam's essentialist natural kinds fit Ellis's scientific essentialism, providing an essentialist realist position. This is realist because the world determines how the groupings of symptoms go together. Haslam's discrete kinds and fuzzy kinds fit Boyd's homeostatic cluster kind position. These could be seen as non-essentialist realist positions. Though not an essentialist position, the world still determines how the groupings of symptoms go together. This is typically thought to make them realist but I will look at arguments which challenge realist understandings of discrete and fuzzy kinds below. Finally, Haslam's practical kinds fit neither Ellis's nor Boyd's account. Psychiatrists, rather

than the world, determine the boundaries of practical kinds. This means they do not fit the two most popular approaches to realism employed by philosophers of psychiatry and therefore practical kinds would be considered an antirealist position.

Essentialist natural kinds are realist and practical kinds are antirealist under the most common approaches in philosophy of psychiatry. In what follows I focus on the more problematic fuzzy and discrete kinds. I consider an argument which would suggest that fuzzy kinds are actually antirealist then I consider another argument which would suggest that both fuzzy and discrete kinds are actually antirealist.

Are psychiatric diagnoses part of the metaphysical structure of the world?

Essentialist kinds and discrete kinds have clear boundaries in comparison to nonessentialist and nondiscrete kinds. Something either is or is not a particular essentialist or discrete kind. Therefore, they seem part of the metaphysical structure of the world since that metaphysical structure seems to determine their boundaries. In contrast, fuzzy kinds do not have discrete boundaries. For any given fuzzy kind, there will be individuals who definitely are that fuzzy kind, individuals who are definitely not that fuzzy kind but there are also individuals who are partly that fuzzy kind. This raises the danger that psychiatrists need to make a decision about who is included in membership of that fuzzy kind. Therefore, the boundaries of the fuzzy kind are not solely determined by the metaphysical structure but also by psychiatrists. This is especially worrying for a realist psychiatry if Haslam is correct to claim that fuzzy kinds are more common in psychiatry than essentialist natural kinds and discrete kinds (Haslam 2014: 23).

A related problem is co-morbidity. This is where an individual is given multiple diagnoses. This situation can be very common. For example, many autistic individuals also meet the criteria for other diagnosis. Around 45 percent of autistic individuals meet clinical criteria for depression and 40 percent meet clinical criteria for anxiety. It is not obviously clear whether this means the individual should be given three different diagnoses or if they should just be diagnosed with autism but the symptoms of autism should also cover the symptoms of depression and anxiety. This suggests the world itself does not determine the boundaries of psychiatric diagnosis but rather psychiatrists must make this decision, which would make fuzzy kinds metaphysically antirealist.

Realists can respond by adopting a pluralistic realism. There are real groupings of behaviors in the mind-independent structure of the world. Therefore, a psychiatric diagnosis which describes that behavior is describing something real, even if there is more than one way psychiatrists can describe that real behavior (Cooper 2007: 50). For example, the symptoms associated with autism are real; additionally, they group together in reality. Despite this, it is still an open question if psychiatrists should just employ autism, should instead employ autism and Asperger's syndrome or use autism, Asperger's syndrome, and even more subtypes.[4] The realist would claim all these options are ways of describing the real world, making each option a realist psychiatric diagnosis. This does, however, raise concerns that psychiatric diagnoses are only real in a very loose sense. The behavior exhibited by patients might be part of the metaphysical structure of the world, but it seems that the opinions of psychiatrists are required to formulate fuzzy kinds, meaning that determining fuzzy kinds involves something more than just the metaphysical structure of the world. Fuzzy kinds seem to sit between essentialist and discrete kinds, which have clear boundaries of symptoms, and practical kinds, which have their boundaries determined by psychiatrists rather than the world. This leaves the metaphysical status of fuzzy kinds unclear.

Are mechanisms part of the metaphysical structure of the world?

Essentialist kinds are realist because they have underlying essentialist mechanisms, that is, a mechanism which is present in every instance of the kind and not present elsewhere. Under this approach the world determines the boundaries of the mechanism and therefore also determines the boundaries of the essentialist kind. Discrete and fuzzy kinds are taken as real because they have homeostatic mechanisms. However, I now discuss the possibility that homeostatic mechanisms are determined not by the world but by the views of psychiatrists. This would in turn mean that discrete and fuzzy kinds are determined by the views of psychiatrists rather than the world and so would not be realist.

Homeostatic mechanisms might be demarcated on the views of psychiatrists rather than the metaphysical structure of the world because diverse causes can produce a homeostatic mechanism. For example, species continue to have relatively similar characteristics due to the homeostatic mechanism of the exchange of genes during mating, however, the exact same genes will not be transmitted in each instance of mating. Therefore, one homeostatic mechanism,

the breeding habits of a particular species, will contain much causal diversity, specifically, different genes being transmitted in any particular instance of two animals of that species mating. Philosophers of psychiatry are rarely clear about the degree of causal diversity that can be present before a homeostatic mechanism stops being demarcated by the metaphysical structure of the world and is instead demarcated by the views of psychiatrists. I shall first discuss the current evidence of causal diversity in psychiatry then discuss the philosophical issues this raises.

Prominent psychiatrists David Kupfer, Michael First, and Darrel Regier write that "not one laboratory marker has been found to be specific in identifying any of the DSM-defined syndromes. Epidemiologic and clinical studies have shown extremely high rates of co-morbidity among the disorders, undermining the hypothesis that the syndromes represent distinct etiologies" (Kupfer et al. 2002: xvii). Philosophically informed psychiatrist Kenneth Kendler writes that "our genes seem neither to have read DSM-IV nor to particularly respect the diagnostic boundaries it established" (Kendler 2010: 1291). Psychiatry has uncovered very few mechanisms responsible for currently employed diagnoses and modern psychiatric evidence suggests there might not be many essentialist or homeostatic mechanisms out there waiting to be found. Rare instances of identifying big, simple causes are unlikely to be replicated (Kendler 2005a: 434). Stronger causes are easier to find, such as genes which are relatively common in individuals with a particular diagnosis being easier to find than genes which occur very rarely within that diagnosis (Kendler 2005b: 1247). Additionally, causation in psychiatry often does not work in the manner of one cause one effect. Often, the same cause can result in many different outcomes while different causes can result in the same outcome.

Are such levels of causal heterogeneity compatible with homeostatic mechanisms or not? Unfortunately, very few philosophers define in detail what exactly constitutes a homeostatic mechanism. Consider that, for example, autism is currently being linked to hundreds of genes. Do hundreds of possible genes, only some of which will be present in any autistic person and many of which can be found in non-autistic people, count as a mechanism? An alternative approach to mechanisms focuses on developmental pathways, whereby lots of different causes (such as the presence of some of those hundreds of genes linked to autism) can set off a process. This process would be the homeostatic mechanism that causes autism. Therefore, a psychiatric diagnosis could have a single underlying homeostatic mechanism even if an immense variety of causal factors determine if that homeostatic mechanism actually is present. However,

current evidence suggests many different causal pathways can produce grouping of symptoms. For example, Sun et al. (2010) found twenty-four possible causal pathways which can lead to schizophrenia. Does this mean schizophrenia has twenty-four different homeostatic cluster kind mechanisms and therefore should be split into twenty four different diagnoses, or does each of those pathways count as part of one bigger homeostatic cluster kind mechanism? Philosophers of psychiatry appealing to cluster kind mechanisms have generally not discussed such matters and those who have discussed this often disagree.

All this raises concerns that the metaphysical structure of reality in psychiatry might not consist of homeostatic mechanisms. Discrete kinds and fuzzy kinds are considered realist because they have real mechanisms; however, if homeostatic mechanisms are things which psychiatrists demarcate, rather than existing in the metaphysical structure of the world, then discrete kinds and fuzzy kinds would not be realist. Either there needs to be substantially stronger causes out there which have not yet been discovered, though there is reason to doubt this possibility, or realist philosophers of psychiatry need develop their position if they wish to avoid metaphysical antirealism.

Epistemic realism and antirealism

The second realist concern is epistemic realism. Scientific realism also requires the "*The Epistemic Thesis*: Mature and predicatively successful scientific theories *are* well confirmed and approximately true of the world" (Psillos 2009: 4, emphasis in original). These are epistemic questions about how much we can know about the metaphysical structure of the world. An epistemic realist believes that scientific theories correctly describes the mind-independent structure. An epistemic realist believes there is something that a particular scientific theory could describe and that it describes it correctly. In contrast, an epistemic antirealist doubts that scientific theories correctly describe the world. An epistemic antirealist might be a metaphysical realist, believing there is something that a particular scientific theory could describe, but doubts that scientific theories accurately describe that part of the world. For example, an epistemic realist in physics might believe the metaphysical structure of the world contains particles and believe that our theories about electrons correctly describe some sub-atomic particles. In contrast, an epistemic antirealist might accept that sub-atomic particles exist but might also believe our theories about electrons do not accurately describe the real sub-atomic particles. It might be

the case that we group different types of sub-atomic particles under the name electrons. In reality, the sub-atomic particles are actually different, having different properties, but physicists mistakenly believe they are all the same sub-atomic particle which we name electrons.[5]

Typically, epistemic realism and epistemic antirealism are theory specific; few philosophers adopt scientific realism regarding all scientific theories or adopt antirealism regarding all scientific theories. Modern scientific theories differ in how speculative they are and how much they are supported by evidence. An epistemic realist about a particular psychiatric diagnosis believes it substantially describes the metaphysical structure of the world; an epistemic antirealist about a particular diagnosis doubts this. As with scientific theories, it is possible to be realist about some psychiatric diagnosis and not others.

Questions of epistemic realism have been long studied within philosophy of science. These are general arguments that can be applied to all sciences. Since psychiatry is typically considered a science I will apply these general arguments to psychiatry. Having done this, I will then look at epistemic issues which are specific to psychiatry.

Epistemic arguments in philosophy of science

The main argument for scientific realism within philosophy of science is known as the no miracles argument. This argument points to the incredible predictive success of scientific theories. Some scientific theories, especially in physics, are capable of making immensely accurate predictions. The argument goes that there are only two possible ways scientific theories are capable of making such predictions: either the theory correctly describes the world or a miracle has occurred. Since miracles are not allowed in philosophy the only explanation is that scientific theories correctly (or very nearly correctly) describe the world. The no miracles argument is rarely applied to psychiatry, including by philosophers of psychiatry. It is not easy to apply to psychiatry since examples of miracle like predictions are rare in psychiatry. It can, however, be applied in a more minimal form. If a particular psychiatric diagnosis were considered successful then it seems reasonable to suppose that the diagnosis is making at least some partially accurate claims about the world. A successful psychiatric diagnosis might be one which significantly helps predict how people with the diagnosis will act and helps with understanding why they do so. However, it is unclear just what degree of success would be required for this argument to work. Additionally, if a psychiatric diagnosis were considered successful this would

only entail that it makes at least some partially accurate claims about the world but the exact degree of accuracy would not be clear.

I now discuss three arguments for epistemic antirealism found in the philosophy of science, starting with the pessimistic meta-induction. Throughout history scientists have believed in theories, often on extremely good evidence, which turned out false, suggesting good evidence is insufficient for belief and our current theories will turn out false. Past and potential future changes suggest modern psychiatric diagnoses are on shaky ground and are likely to change. This likelihood of being abandoned undermines reliable belief in them. Arguments similar to these are common in philosophy of psychiatry, with regular complaints that psychiatric diagnoses keep changing. It is unclear why psychiatric diagnosis keep being revised if they are already correct. There are concerns that these changes are driven by the arbitrary views of psychiatrists rather than scientific evidence. Additionally, advances in genetics and neuroscience may mean we abandon major diagnoses like schizophrenia and manic depression. Past and potential future changes are often taken to suggest current psychiatric diagnoses do not merit belief.

Another argument from philosophy of science is underdetermination. This is where two different theories can equally well explain the evidence; therefore there is no justification for assigning belief to one theory rather than the other. If more than one theory is equally compatible with the evidence then scientists cannot employ evidence to decide which theory is correct. Related arguments in philosophy of psychiatry are common, with the suggestion that current psychiatric diagnoses are just one way of classifying mentally ill people. Murphy writes, "What evidence do we have that current practice in psychiatry produces good outcomes, outcomes that couldn't be matched by alternative nosologies that start from a completely different set of premises?" (Murphy 2006: 10). The possibility of constructing alternative psychiatric diagnoses is taken to suggest we lack reason to believe current psychiatric diagnoses are the correct ones.

A third philosophy of science argument concerns the theory-laden nature of evidence. This is where evidence employed to construct scientific theories itself partly depends upon scientific theories, potentially mistaken ones, which can result in substantial negative epistemic consequences. This also has parallels in philosophy of psychiatry, it being commonly argued that psychiatric diagnoses are overly reliant upon false theories. There are many competing theoretical approaches to mental illness, such as psychoanalysis, cognitive psychology, neuroscience, and genetics. Not only are we unsure which theory is correct, potentially many of them are false. If so, theories might cause psychiatrists to

make false observations, theory influencing and thus distorting the observations used to form diagnoses. Reliance upon false theories undermines belief in existing psychiatric diagnoses.

I believe that these arguments are best judged by historical studies. The strength of the pessimistic meta-induction is based upon establishing how greatly a psychiatric diagnosis undergo change in criteria across their history. Underdetermination arguments are best judged by comparing modern diagnosis to actual historical diagnoses which have stood the test of being applied to real patients rather than to undeveloped hypothetical alternatives. Theory-laden nature of evidence arguments are also best judged by studying if psychiatrists in the past, who often held very different theories to modern psychiatrists, reported different symptoms or clustered symptoms together in a different manner.

Epistemic arguments in philosophy of psychiatry

I now consider epistemic arguments in philosophy of psychiatry. These are arguments discussed by philosophers of psychiatry which deal with issues specific to psychiatry rather than being arguments applicable to all sciences.

There are many epistemic issues over neuroscience. For example, there are concerns that the dominant causal theory of schizophrenia since the 1970s, the dopamine hypothesis, is mistaken and should be abandoned. These issues are important because neuroscience is often taken as the most likely place to find the mechanisms underlying essentialist, discrete, and fuzzy kinds. There is debate over whether neuroscience is ready for finding mechanisms underlying psychiatric illness. Some philosophers think neuroscience is ready whereas other consider neuroscience to be insufficiently developed. Additionally, even when sufficiently developed, employing neuroscience to understand the multiplicity of causal factors will be very difficult, raising epistemic concerns that mechanisms are incorrectly described or one mechanism is unrealistically emphasized over other mechanisms.

There are many epistemic concerns over how well psychiatrists describe the symptoms of patients. Phenomenologists are concerned that psychiatrists use relatively basic understanding of symptoms which does not convey the complexity of subjective experience. Also, there are often relatively weak links between symptoms and underlying neurology, since symptoms often involves non-neuroscientific factors like social environments. This creates epistemic issues over knowing the degree to which a neuroscientific or nonneuroscientific understanding is required.

There are also epistemic issues related to nonscientific interests influencing scientific research. For instance, the insurance industry and drug manufacturers could influence psychiatric researchers to consciously or unconsciously interpret results to suit their interests rather than just describe the world.

Some philosophers have believed that patients will respond to different drugs because they have different underlying mechanisms, therefore drug responses help establish which patients share underlying causes (Tsou 2012: 546). However, drugs typically only affect individual symptoms and typically affect individuals with different psychiatric diagnosis, making drugs epistemologically unreliable for demarcating psychiatric diagnosis.

There are epistemic concerns that psychiatric diagnosis distracts psychiatrist's attention away from other aspects of the patient. There are concerns that many important aspects which influence how the patients behave, such as their upbringing, their socioeconomic environment, and their personality are largely ignored. These factors could significantly cause mental distress but they are not typically considered medical factors. For example, if an individual is depressed because they have poor housing then the obvious solution would be to provide them with better housing rather than send them to a psychiatrist. Epistemic issues arise when such factors are ignored due to focusing excessively upon their psychiatric diagnosis.

Overall, philosophers of psychiatry are on balance epistemic antirealists. They are generally concerned that currently employed psychiatric diagnoses are flawed and could be improved to better describe reality. They vary in the degree to which they believe psychiatric diagnosis need modifying, some calling for widespread changes and others calling for more minor changes.

Conclusion

Psychiatrists want psychiatric diagnoses to describe something real, rather than something in the minds of psychiatrists. This would help ensure the diagnosis actually describes what the patient is like and help the development of better treatments.

Realism has two components, the metaphysical and the epistemic. Most philosophers of psychiatry are metaphysical realists, believing there is a mind-independent structure of natural kinds for psychiatric diagnosis to describe. Natural kinds are typically taken as having boundaries of symptoms and mechanisms which are demarcated by the world. However, if, as I have argued, psychiatrists

need make choices when drawing the boundaries of symptoms and if the immense causal diversity underlying psychiatry mean mechanisms are demarcated by psychiatrists then notions of realist psychiatric diagnosis are challenged. Finally, many philosophers of psychiatry are epistemic antirealists, believing most currently employed psychiatric diagnoses fail to accurately describe the world.

The realism debate over psychiatric diagnosis has important consequences. If realism is justified, then the world largely or entirely determines what the correct psychiatric diagnoses are. In contrast, if antirealism is justified then psychiatrists need decide which psychiatric diagnosis to employ since there is no one correct set of psychiatric diagnosis. If so, then psychiatrists need consider and debate what they want psychiatric diagnosis to be like. For example, they need decide whether they want to employ lots of diagnosis, each of which provides a quite specific description of patients, or if they want to employ a smaller number of diagnosis, each of which is applicable to a wider range of individuals. Thus antirealism over psychiatric diagnosis entails that there should be a debate over how to formulate psychiatric diagnosis. This debate need not take place if metaphysical and epistemic realism over psychiatric diagnosis were justified, meaning psychiatry had found the one correct set of psychiatric diagnosis.

Acknowledgments

I would like to thank Rachel Cooper, Dan Degerman, Hane Maung, Moujan Mirdamadi, Jane Elouise Pitt, and the editors of this volume, who read and offered helpful criticism of earlier versions of this chapter.

Notes

1 Philosophers of psychiatry typically see psychiatric diagnoses as entities rather than scientific theories. Some scientific realists, known as entity realists, also see scientific realism as focusing on entities rather than theories. However, most scientific realist typically focus upon scientific theories because they believe entities can only be established through scientific theories. I believe this is the case with psychiatric diagnoses. Consequently I shall focus upon typical notions of scientific realism and its emphasis on theories.

2 Note that historians and historically inclined philosophers often disagree over the degree to which abandoned scientific theories and abandoned psychiatric diagnosis actually were mistaken; these are examples of problematic cases.

3 There is also "*The Semantic Thesis*: Scientific theories should be taken at face value.… they are capable of being true or false," Psillos (2009: 4) emphasis original. I shall not discuss this because semantic anti.-realism is generally rejected by modern philosophy of science. "Semantic realism is not contested anymore," Psillos (2009): *5* with even leading antirealists philosophers of science being semantic realists.

4 This example reflects the difference between *DSM*-IV and *DSM*-5. *DSM*-IV included the diagnoses of autism and Asperger's syndrome, whereas *DSM*-5 merged Asperger's syndrome into autism.

5 Note that it is possible to be a metaphysical antirealist but also an epistemic realist. A philosopher might think there is no one true structure for science to describe yet still thinks scientific theories can be assessed for the level of belief they deserve. For example, an epistemic realist might think simple theories are preferable to complicated theories and so use simplicity as a standard to judge which theories merit belief. They would only be epistemic realist over simple theories even though, by being metaphysical antirealists, they are not willing to claim that the nature of reality itself is simple.

References

Beebee, H. and Sabbarton-Leary N. (2010), Are Psychiatric Kinds "Real"?, *The European Journal of Analytical Philosophy*, 6:11–27.

Boyd, R. (1991), Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds, *Philosophical Studies*, 61:127–148.

Cooper, R. (2007), *Psychiatry and the Philosophy of Science*, Stocksfield: Acumen.

Ellis, B. (2001), *Scientific Essentialism*, Cambridge: Cambridge University Press.

Haslam, N. (2002), Kinds of Kinds: A Conceptual Taxonomy of Psychiatry Categories, *Philosophy, Psychiatry and Psychology*, 9:203–217.

Haslam, N. (2014), Natural Kinds in Psychiatry: Conceptually Implausible, Empirically Questionable, and Stigmatizing. In *Classifying Psychopathology*. H. Kincaid, J. A. Sullivan, eds. Cambridge, MA: MIT Press.

Kendell, R., Jablensky, A. (2003), Distinguishing between the Validity and Utility of Psychiatric Diagnosis, *American Journal of Psychiatry*, 35:9.

Kendler, K. S. (2005a), "A Gene for … ": The Nature of Gene Action in Psychiatric Disorders, *American Journal of Psychiatry*, 162:1247.

Kendler, K. S. (2005b), Towards a Philosophical Structure for Psychiatry, *American Journal of Psychiatry*, 162:434.

Kendler, K. S. (2010), Advances in Our Understanding of Genetic Risk Factors for Autism Spectrum Disorders, *American Journal of Psychiatry*, 167:1291.

Kendlar, K. S., Zachar, P., and Craver, C. (2011), What Kinds of Things Are Psychiatric Disorders?, *Psychological Medicine*, 41:1143–1150.

Kupfer, D. J., First, M. B., and Regier, D. A. (2002), *A Research Agenda for DSM-V*, Washington, DC: American Psychiatric Association.

Laudan, L. (1981), A Confutation of Convergent Realism, *Philosophy of Science*, 48:19–49.

Murphy, D. (2006), *Psychiatry in the Scientific Image*, Cambridge, MA: MIT Press.

Murphy, D. (2014), Natural Kinds in Folk Psychology and in Psychiatry, pp. 105–122. In *Classifying Psychopathology*. H. Kincaid and J. A. Sullivan, eds. Cambridge, MA: MIT Press.

Psillos, S. (2009), *Knowing the Structure of Nature: Essays on Realism and Explanation*, Basingstoke: Palgrave Macmillan.

Sun, J., Jia, P., Fanous, A. H., Van Den Oord, E., Chen, X., Riley, B. P., Amdur, R. L, Kendler, K. S., and Zhao, Z. (2010), Schizophrenia Gene Networks and Pathways and Their Application for Novel Candidate Gene Selection, *Plos One*, 5:e11351.

Tsou, J. Y. (2016), Natural Kinds, Psychiatric Classification and the History of the DSM, *History of Psychiatry*, 27:406–424.

Tsou, J. Y. (2012), Intervention, Causal Reasoning, and the Neurobiology of Mental Disorders: Pharmacological Drugs as Experimental Instruments, *Studies in History and Philosophy of Biological and Biomedical Sciences*, 3:542–551.

social aspects, for example, regarding the funding of research, the formulation of criteria for epistemic success and expertise, or potential values and background assumptions informing research questions and methodologies. Another aspect to consider is psychiatry's often controversial status, in scientific terms as well as in its public reception. The complexities of its subject matter as well as of the societal demands it is confronted with pose special challenges for the social organization of psychiatric knowledge-production and -application, making it a highly interesting and important case for social epistemologists.

In the following, I will give an overview of the different questions discussed in social epistemology and mention connections to psychiatry throughout. The section "Testimony" starts with the issue of testimonial knowledge; that is, knowledge based on information I get from others. The section "Agency" introduces questions that concern the potential social character of the subjects of knowledge and the production of a great deal of (psychiatric) knowledge by groups rather than individuals. The section "Objectivity" constitutes the main part of the chapter and deals with the social epistemology of science. After summarizing the historical background of its development, I will sketch the various roles social values can play in science Afterwards I introduce different models of how to integrate scientific objectivity with the social aspects of research and show how these relate to current issues in (philosophy of) psychiatry, such as psychiatric classification and the debates surrounding the latest revision of the *Diagnostic and Statistical Manual* (*DSM*-5; APA 2013).

Testimony

A lively discussion has developed on the issue of testimonial knowledge. Traditional epistemology has discussed sources of knowledge for individuals (such as perception, introspection, memory, or reasoning) and their reliability in great detail. However, many of our beliefs are based on information we get from others, creating the question whether to add this to the list of sources. David Hume, founding the historically dominant view of testimonial reductionism, argued that testimony can only be considered a source of knowledge in a derived sense: while we often are justified in believing what we are told by others, this justification is based on our experience of former instances of testimony being verifiable by our own cognition and perception (Hume 1748; cf. Reid 1764 for a contemporary counter position). By contrast, anti-reductionists think that we are prima facie justified in believing the testimony of others.

Coady (1992) maintains that refusing testimony as a genuine social source of knowledge leaves us with a too small number of justified beliefs. For example, he argues that even beliefs based on individual perceptions presume socially acquired background knowledge that informs the conceptual representation and interpretation of our perceptions. It is important, however, that anti-reductionism does not imply that one can justifiably believe whatever others say; rather, it holds that we are justified in believing the testimony of others in cases where we do not have any reasons against believing it (Burge 1993). The current discussion on testimony mostly takes place in-between these poles and concerns questions such as what kind of reasons (e.g., only epistemic or also moral) can speak for or against testimonial credibility, how demanding the conditions for believing or refusing testimony have to be, and whether and how a middle ground can be established (e.g., Fricker 1994; Goldberg 2010; Lackey 2008).

In addition to such general questions, there are discussions on more specific issues, such as *expertise* or *testimonial justice*, which are of high relevance to philosophy of psychiatry. In many societal contexts, informed decision-making makes it necessary to rely on the testimony of experts, especially scientific experts. For example, this might concern the use of forensic experts in court trials, the reliance on climate scientists in environmental policy making, or simply a sick person's doctor's visit. The need for such expert testimony stems from a knowledge gap between lay persons and experts, which can make relying on others' testimony the best epistemic strategy. However, it is precisely this gap that also leads to difficult questions, such as how lay persons can identify trustworthy experts or how to deal with dissenting experts (e.g., Goldman 2001; Hardwig 1985; Selinger and Crease 2006).

Such *dissent between experts* not only poses the question whose testimony to prefer on what grounds, but a continued lack of consensus between experts can also lead to a more general distrust in scientific expertise. Sometimes, such dissent is deliberately manufactured or enforced, as examples from tobacco research or climate science show (Oreskes and Conway 2010).[2] In psychiatry, expert disagreements often stem from an actual lack of consensus and knowledge, as the discipline is dealing with a highly complex subject matter and comprises work under different paradigms (e.g., biological psychiatry or psychoanalysis). This expert dissent is often accompanied by public disputes about whether conditions such as ADHD are real disorders at all, or are better understood as within a range of normal variation of human behavior. Such issues are particularly pressing in psychiatry because of the social and ethical implications

of diagnoses of mental illness. While a diagnosis can of course be beneficial in so far as it grants access to health care, treatment, and maybe reimbursement, there are widespread worries about an exaggerated medicalization and, in particular, pharmaceuticalization in the psychiatric realm (cf., e.g., Conrad and Potter 2000; Lane 2007). This is closely connected to skepticism and distrust toward expert testimony because of financial ties to pharmaceutical companies (cf., e.g., Cosgrove et al. 2006; Cosgrove and Krimsky 2012). An important question for philosophy of psychiatry is thus how to reconcile expert dissent and public credibility (see also the section "Objectivity").

The second issue of particular relevance to philosophy of psychiatry revolves around the idea of *testimonial injustice*, introduced by Fricker (2007) as a case of epistemic injustice. It refers to situations where a speaker's testimony receives less trust and is ascribed lower credibility than it deserves due to prejudices against the speaker's social group. Since a diagnosis of mental disorder (or even somatic disease) is still often accompanied by stigmatization, an interesting new research question for philosophy of psychiatry is to explore the connection between such stigmatization and testimonial injustice (e.g., Carel and Kidd 2014; Crichton et al. 2017). To what extent and under what conditions are mentally ill patients suffering from such unjustified distrust in their testimony? Insofar such testimonial injustice regarding mentally ill patients exists, the next question is how to remedy it. Miranda Fricker's approach builds on individuals' ability to critically reflect on their bias and prejudices. As this ability is often limited, others have argued that it is necessary to think about how epistemic and testimonial justice can be supported or hindered by social structures and institutions (e.g., Anderson 2012)—thereby opening another important application for social epistemology.

Agency

Apart from social sources of knowledge (and ignorance), social epistemology is concerned with issues regarding the agents of knowledge. Are these always individuals, or can social groups be considered as epistemic or doxastic agents?[3] While we do often attribute beliefs to groups (e.g., in ascribing certain positions to a political party or committee), a reductionist account of group knowledge would consider this as metaphorical or imprecise speech, arguing that only individuals can actually hold beliefs. Opposed to this, others have argued that groups are in fact the primary knowers at least in some contexts (such

as empirical science), that individuals can obtain (scientific) knowledge only in virtue of their being a member of a (scientific) community and taking part in group interactions, or that the very standards of knowledge depend on the respective group (e.g., Kusch 2004; Longino 2002; Nelson 1990).

In any case, questions arise regarding the conditions under which beliefs can be correctly ascribed to group agents. *Summative models* assert that a group can be said to hold a certain belief under the condition that most or all of its members hold this belief. Some authors argue that this condition is too minimal and the ascription of group beliefs comes with further requirements, such as the group members expressing a willingness to convey that belief as the official position of their group (Gilbert 1989; cf. also Goldman and Blanchard 2016). Another problem for summativism is that we sometimes ascribe beliefs or knowledge to groups instead of individuals precisely because this group knowledge exceeds shared beliefs of the group members. Such cases are represented by *distributive models* of cognitive agency (Bird 2014), which are often concerned with groups pursuing epistemic goals that due to their complexity require a division of cognitive labor, as a high degree of complexity can make it impossible for a single person to know about and thus believe all relevant propositions.[4]

Exemplary of such epistemic endeavors relying on a division of labor is scientific research, especially in cases where research is done by large groups or is made possible only by knowledge-intensive technology. The generation of knowledge by such groups does not require that all participants share, or are even aware of, all of the relevant beliefs; instead, it requires that there are mechanisms of distributing subtasks to participants with special expertise. Especially interesting in this regard are cases of interdisciplinary research, as this may come not only with different areas of expertise complementing each other, but might also be characterized by differences in background assumptions or standards of epistemic and methodological quality (e.g., Andersen and Wagenknecht 2013).

While general ontological questions regarding group agency may not be of special relevance to philosophy of psychiatry, the epistemology of research groups certainly is. Moreover, psychiatric research requires interdisciplinary collaboration, and psychiatry is still characterized by the existence of different paradigms in research and treatment. Successful epistemic collaboration in such complex contexts raises the issue of how to best coordinate it and how to deal with the given pluralism in the field, especially in the face of shared tasks such as psychiatric classification. Fulfilling such an applied task moreover often requires the building of an expert consensus despite the presence of

scientific dissent or uncertainty. The procedures of building such a consensus are also a very fruitful socio-epistemological topic, which connects with social scientific work on potential pitfalls of group deliberations, such as groupthink or the marginalization of dissenting voices (Solomon [2015a] provides a rich account of the procedures and functions of such processes in NIH consensus conferences; Solomon and Kendler [2016] apply this to the role of expert consensus in psychiatric classification).

Objectivity

Development of the social epistemology of science

Science is the paradigmatic epistemic enterprise, being aimed at the systematic production of knowledge. It is also distinctively social: its very existence depends on conditions such as the availability of funding and of certain educational and academic institutions, as well as the political freedom for critique, discussion, and dissemination of research results. Scientists themselves are social beings situated in a particular societal, economic, political, and cultural context impacting their background assumptions and motivations. A great deal of modern science is not done by individual scientists, but by research groups of various sizes collaborating and dividing epistemic labor. Nevertheless, philosophy of science (like epistemology) has for a long time excluded the social aspects of research from its considerations. The reason for this can be found in a normative standard of *good science,* or *objectivity,* as requiring the elimination of social impacts (especially value-judgments) on scientific results. This elimination was, moreover, thought to be achievable by adhering to methods that could, at least in principle, be implemented by individuals.

While it was always clear that science does not happen in a social vacuum, philosophy of science, from its beginning as an institutionalized sub-discipline in the early twentieth century, was founded upon a certain framework that ensured the epistemic realm as its independent subject. Reichenbach (1938) explicates this independence in terms of a distinction between the contexts of *discovery* and *justification* (now often complemented by *application*; cf. also Schickore and Steinle [2006] on the history and validity of the context distinction). This distinction is connected to a hypothetico-deductive model of scientific testing: according to this, ideas and hypotheses are generated in a process that can be highly creative as well as influenced by political values or prejudices. In a second step, however, these hypotheses are tested empirically. The results of this test

are justified only if they are independent from all the social aspects and value-judgments figuring in the context of discovery. These results can moreover be put to practical use in the context of application, which requires decisions that are often informed by ethical, political, or economic values and interests. Again, these supposedly do not affect the context of justification; whether we decide to take certain measures against global warming, for example, does not have any epistemological bearing on whether hypotheses on the causes of climate change are justified. According to this view, the internal epistemic aspects of scientific research are potentially and ideally independent from its external social aspects (cf. e.g., Koertge 2000). Consequently, theoretical philosophy of science is concerned exclusively with epistemology and thus the context of justification, whereas the proper subject matter of any sociological or historical science studies is to be found in the contexts of discovery or application.

This alleged independence of epistemic and social aspects of science has, however, been confronted with serious problems. On the one side, there are systematic points against the overly simplistic model of justification and scientific testing it is built on. Important here are arguments based on the thesis of an *underdetermination* of theory by evidence, which holds that empirical evidence is (often or in principle) insufficient to unequivocally decide whether to accept or reject a hypothesis. *Holistic underdetermination* refers to the idea that we cannot confront individual hypotheses directly with empirical data, since these hypotheses are always located in a complex web of other assumptions (e.g., about the measuring methods and instruments). Even if the data does not fit the hypothesis under test, it might thus be that the mistake lies somewhere else and the hypothesis can be saved by rebuilding other parts of our theoretical web or changing the measuring instrument. *Contrastive underdetermination* consists in the idea that a fit between empirical data and hypothesis cannot prove the latter either, as the same data might also (in principle) be explainable by other theories (Duhem 1906; Quine 1951).[5] On the other side, following the work of Thomas Kuhn and others, a greater concern for actual scientific practice in its respective historical and social context has cast doubt on whether science ever achieves this independence, as well as on the possibility of ahistorical standards of justification (Kuhn 1962).

These developments paved the way for an increase in historical and sociological studies of science that no longer restricted themselves to discovery and application but extended their concern to the very contents and standards of scientific knowledge. Yet, much of this early work in science studies as well as philosophy was still built on the assumption of a fundamental difference

between the epistemic and the social. Therefore, the examination of social impacts on justification was often considered as debunking science's claim to objectivity by postmodernists, social constructivists, or authors belonging to the Strong Programme in Sociology of Knowledge (cf. Brown 2001; Hess 1997). By contrast, the social epistemology of science aims to integrate the social and epistemic aspects and examines under what specific conditions social influences on research can support or compromise epistemic success.[6]

Extending the scope of philosophy of science

Social epistemologists do, however, differ considerably regarding the question of how deep the impact of the social, especially of social values, is or should be. On the more conservative end of the social epistemologist spectrum are accounts that continue to consider objectivity as presuming the value-freedom of results, yet differ from mainstream philosophy of science in no longer conceiving of objectivity as ensured by scientific methods that are in principle employable by individuals. Instead, they argue that objectivity and epistemic success are promoted by certain social mechanisms and institutional measures. For example, the early work of Philip Kitcher has reflected on issues such as an efficient division of epistemic labor or how the reward system of science and its peculiar focus on priority enable scientific progress (Kitcher 1993; cf. also Goldman 1999).

Kitcher's later work (2001) is exemplary of another important line of thought in the social epistemology of science, namely, to connect normative philosophy of science and ideals of democratic deliberation. Here, he offers an account of how to organize the direction of research, thereby extending the scope of philosophy of science to the context of discovery. His ideal of *well-ordered science* (WOS) is based on a rejection of a "myth of purity" regarding scientific agenda-setting and proposes to democratize it. This rejection is directed against the premise that science had one main goal, such as the pursuit of truth. According to Kitcher, such a goal is far too general to serve as a guide here: science does not pursue all truths, but *significant* ones. The determination of significance, however, depends on more specific goals connected to the pursuit of certain problems (instead of others) and can thus be value-laden (Kitcher 2001: 63–93).

Ideally for Kitcher, decisions about the selection of problems (as well as on ethical constraints on methods or the application of results) would represent a consensus among representatives of diverse societal groups and interests. These representatives are conceived of as ideal deliberators, who—tutored by impartial experts—are able to understand and process information about the state of art

concerning different research problems as well as to respect the interests of other parties just as much as their own.

While this ideal may not be achievable, Kitcher takes it to be able to show us when we go wrong by departing too far from it (2001, 117–137). For instance, Kitcher and Reiss (2010) argue that current biomedical research deviates sharply from well-ordered science's aim of promoting the common good. First, agenda-setting in medical (including psychiatric) research is too often aimed at economic profit rather than at providing treatments where most needed, since so much of it is being done by pharmaceutical companies. These companies (naturally) invest in research areas that constitute big markets (e.g., remedies for hair loss, obesity, or erectile dysfunction), rather than allocating research funds to issues such as malaria or tuberculosis, which come with a higher level of global disease burden but promise less financial payoff. Regarding psychiatry, it has been argued that private funding has led to an exaggerated focus on pharmaceutical solutions for mental illness at the expense of research on their ecological and psychological conditions (Musschenga et al. 2010). To decrease this deviation from WOS, Kitcher and Reiss consider it important to discuss sociopolitical measures, such as the amount and distribution of public funding as well as the incentives provided by certain regimes of intellectual property and patenting.

Secondly, and related to the point on agenda-setting, there is a lack of impartial experts in biomedical research, as, for example, research results often display a systematic bias in favor of the respective companies' interests (e.g., Lundh et al. 2015). To improve objectivity, many authors have moved beyond the individual level in proposing institutional counteractions such as the registration of all clinical trials and publication of all respective data, the disclosure of financial ties of experts, or the independent testing of pharmaceutical drugs (e.g., Brown 2008; Jukola 2015; Kitcher and Reiss 2010; Sismondo and Doucet 2010).

Science and values

Bias in the results of pharmaceutical research is only one example where economic interests or value-judgments transcend agenda-setting and enter into theory assessment. There is a substantive amount of literature on the relation of objectivity and value-freedom in social epistemology, revolving around a number of arguments for legitimate or inescapable roles of values in science.

For instance, it has been argued that some terms, such as "economic productivity" or "rape" cannot be reduced to descriptive criteria unaffected

by values and interests (Dupré 2007). A good example from philosophy of psychiatry is the notion of mental disorder. Generally, there is an extensive debate on whether it is possible to give a value-free definition of "health" and "disease." Naturalists argue that this can be done in terms of criteria for normal functioning (Boorse 1977), while normativists hold that the determination of a range of normal function also often involves value-judgments (e.g., Cooper 2002; Nordenfeldt 1987). The most successful account of *mental disorder* explicates it in terms of a "harmful dysfunction." According to this, for a condition to be a mental disorder it is necessary that it involves an evolutionary dysfunction—some mental mechanisms fail to bring about the effects that were responsible for their natural selection. Such a dysfunction is, however, only sufficient for disorder status if it is also evaluated as harmful in a given social context (Wakefield 1992). Accordingly, schizophrenia (or, rather, schizophrenia spectrum disorders) may cross-culturally represent a dysfunction, but not a disorder when viewed as a sign of spiritual power, for example.[7]

If we accept that values play a role in the determination of what is a mental disorder, this leads to questions about how to deal with such issues. Questions of undue medicalization and pathologization are abundant in psychiatry, especially with regard to classifications such as the *DSM* or *ICD*. This represents an important connection to accounts in social epistemology that aim to specify ways for a fruitful management of value influences in science.

Besides such semantic value-ladenness, others have argued that values can impact background assumptions relevant to the interpretation of data (Longino 1990), assessments of the pursuitworthiness of theories and comparative theory choice (Bueter 2015; Elliott and MacKaughan 2009; Ohkrulik 1994), or the setting of standards of evidence. Regarding the latter, for example, the argument from *inductive risk* has been very influential. It concerns research with foreseeable applications, for example clinical trials on the efficacy of a certain antidepressant. If assessed as effective, this drug will likely be used in the treatment of psychiatric patients. Now in assessing the drug, researchers are facing two kinds of error possibilities—false positive results, assuming the drug is effective when it is not, or false negative results, judging the drug to be ineffective when it would in fact be helpful to many people. The argument from inductive risks holds that balancing these kinds of error possibilities requires weighing the associated practical consequences in ethical terms. Is the condition under concern so devastating that one should try to avoid false negative results, even if that means lowering the standards of evidence required for drug approval? Or should one aim to avoid false positives in order to avoid the costs of prescribing

patients a medication that does not work, for example, by increasing the level of significance? While it would obviously be best to reduce both kinds of error risks, scientific research is in principle fallible, making it necessary to think about how sure one needs to be before putting hypotheses into practice, and how to strike the balance between the different associated risks. In affecting such decisions, value judgments are said by many to legitimately impact epistemic standards (e.g., Douglas 2000; Wilholt 2009).

Since much psychiatric research is aimed at successful application and treatment, these questions are especially pertinent here. They also connect to the aforementioned issues of commercialized research, as the risks involved, and the preferable balance of their probabilities, often differ from the standpoint of producers (who want to avoid losing money due to false negative results) and consumers (who might be more interested in avoiding false positives, especially if there are other treatment options available).

Socializing objectivity and the value of pluralism

Taking these arguments about values in science into account, some philosophers have argued for a reconception of objectivity that does not presume value-freedom. Prominently, Helen Longino (1990) has argued for a social view of objectivity as largely based on features of the scientific community. *Social objectivity* aims to make value-impacts in science visible through a process of mutual criticism. The fertility of this process and its ability to generate trustworthy results depends on a number of community-related conditions: First, it requires the existence of recognized avenues for criticisms, such as peer-reviewed journals or conferences. Secondly, there needs to be uptake of such criticism; the scientific community has to respond to dissent. Thirdly, this criticism needs a shared basis in the form of public standards of theory assessment for a given research field (e.g., criteria of evidence-based medicine). These standards are, however, also the result of a critical deliberation and potentially impacted by social values (cf. Longino 1996). Fourthly, the critical process has to be characterized by an equality in intellectual authority of all qualified participants (Longino 1990: 62–83).

This last point is especially important since values and bias often go unnoticed if shared by all members of a scientific community; the detection of value-impacts is thus promoted by the presence of a plurality of theoretical perspectives as well as diversity in the participants' background. For example, a too homogeneous context might come with value-laden assumptions about homosexuality that become invisibly entrenched in its treatment as a mental disorder, or shared

normative ideals of academic success may impact the diagnostic criteria of ADHD (Hawthorne 2010). However, even adherence to the conditions of social objectivity cannot guarantee value-free research results, as these might still be affected by value-laden standards guiding theory assessment or by shared value-laden background assumptions in the community. Longino's account is therefore not only social, but also procedural in that it ascribes objectivity primarily to the process and context of knowledge production.

By contrast, Solomon (2001) develops a more consequentialist ideal of *social empiricism* that focuses on the results of these processes. Solomon introduces the notion of *decision vectors* that guide theory assessment and are of two kinds: empirical decision vectors, for example, salience and availability of data, and non-empirical ones. The latter include not only value judgments but also other factors such as personality traits (e.g., conservativeness or competitiveness), institutional settings, or financial incentives. Rather than focusing on individuals, Solomon argues that scientific rationality is a function of the distribution of decision vectors—empirical decision vectors have to be distributed in proportion to the empirical success of theories, while non-empirical decision vectors need to be present in equal numbers with regard to competing theories, so that they balance each other out.

These discussions about social conditions for objectivity are highly relevant to psychiatry, especially if one accepts the claim that value-judgments play a significant role in research on psychopathologies or the determination and (notoriously controversial) classification of mental disorders (cf. Sadler 2005; Sisti et al. 2013). For example, Ross (2002) argues that Longino's account presents a fruitful approach to questions surrounding decision-making in the *DSM*. Such an application of social epistemological models of objectivity, however, requires further work about their concrete implementation in the realm of psychiatry, and especially psychiatric taxonomy. For example, is the process of *DSM*-revisions currently organized in a way that satisfies the requirement of public avenues of criticism, or is there a lack of transparency? Are the *DSM*-workgroups sufficiently responsive to criticism, and how can this be ensured? What are the specific standards that should guide taxonomic-decision-making, and should they, for example, privilege antecedent or predictive validators?[8]

Of special interest in this regard is also Longino's fourth condition of an intellectual equality of authority of all qualified participants, as it leads to the question of whom to consider qualified to take part in the process. For instance, the revision process has been criticized for suffering from a lack of disciplinary diversity and overly privileging (biological) psychiatry above other perspectives

from psychology or pediatrics (e.g., Sisti and Johnson 2015). Sisti and Johnson also argue that the *DSM* is not an exclusively scientific instrument but has many practical implementations, carrying implications for numerous constituents that should also be represented in the revision process (cf. also Tekin 2014). Of particular concern here is whether, and to what extent, patients or patient advocates should be treated as qualified participants. For example, Porter (2013) argues in favor of increased patient contributions on the basis of an understanding of psychiatric nosology as value-laden; Sadler and Fulford (2004) similarly hold an increased integration of patient voices to be an improvement in ethical as well as epistemic hindsight. Indeed, the *DSM*-5 had taken steps in this direction, for example, by publishing proposals for revisions of diagnostic criteria online at several occasions and asking for public comments (Regier et al. 2010; Stein and Phillipps 2013). More work is required here on different models of patient integration (e.g., via online commentaries or the integration of patient advocates into workgroups or review committees) and how they can promote or hinder (social) objectivity (cf. also Bueter, 2018).

Another issue that has recently been raised by Miriam Solomon is that while the integration of a plurality of perspectives is often fruitful, it can also sometimes be detrimental. In particular, she argues that the dissent often resulting from such a pluralism can be bad in applied research such as the medical disciplines, as dissent here undermines medical authority and public trust (Solomon 2015b). What this shows is the need for future research into the benefits and harms of different forms of (organizing) pluralism in specific contexts such as psychiatric classification. At this point, the social epistemology of psychiatric science and classification moreover connects with issues regarding the sources and agents of knowledge: How can we determine the trustworthiness of lay representatives and their testimony? Is their inclusion into *DSM*-revisions a matter of epistemic justice? How to deal with dissent among not only experts, but also different lay viewpoints as well as lay persons and experts? How can communication in such complex groups be organized efficiently?

Conclusion

Social epistemology is a thriving philosophical research field that has much to contribute to, and in turn to learn from, psychiatry and philosophy of psychiatry. In particular, this holds for the relation between the social epistemology of science and issues connected to psychiatry as a scientific discipline that is

dealing with a complex subject matter in a very specific social context. This challenging situation has led to considerable controversies and debates—from the anti-psychiatry movement of the 1970s to the current debates on psychiatric classification, medicalization, and the role of pharmaceutical lobbying. Many of these debates tend to focus on issues such as the "reality" of mental disorders, the validity of diagnostic categories, or the truthfulness of research results. A socio-epistemological approach can complement these discussions from a procedural point of view that thinks about the harms and benefits of certain institutional structures and incentives, about conditions of credibility and trust as well as about an adequate representation of different points of view and the (legitimate and illegitimate) roles of values in psychiatry.

Notes

1 Goldman and Blanchard (2016) give a more detailed introduction to genuinely epistemological discussions. They develop a threefold taxonomy of social epistemology that I will roughly follow here. It should be mentioned, however, that their article does not mention the important role of feminist epistemology in the development of the social epistemology of science; cf. Anderson (2015); Grasswick (2016) on this intersection.

2 An important line of research in the social epistemology of science consists in agnotology, which deals with such questions as the fabrication of dissent and different sources of ignorance (Proctor and Schiebinger 2006; cf. also Biddle and Leuschner 2015 on epistemically detrimental versus beneficial dissent in science).

3 "Epistemic agency" here refers to any kind of cognitive or knowledge-related acts; "doxastic agency" refers more specifically to the act or state of holding a belief.

4 Cases where not all group members share a set of beliefs moreover lead to problems concerning the aggregation of beliefs. For example, it can be the case that while different positions in a group each consist of coherent subsets of beliefs, a simple summation of individual beliefs contained in these subsets might lead to the ascription of an incoherent position to the group (List and Pettit 2002).

5 The underdetermination thesis is a complex and controversial issue; it should not be taken to show that scientific objectivity and rationality are illusions (as it sometimes is), since there can be other factors than empirical evidence figuring in the rational evaluation of theories, for example, explanatory scope or coherence with background knowledge (cf. Laudan 1990; Potter 1996 on different interpretations of underdetermination).

6 Throughout, these developments have been connected to and shaped by discussions in feminist philosophy of science. Harding (1986) usefully distinguishes between

three broad kinds: *Feminist empiricism* criticizes androcentric or sexist bias in science while adhering to given standards of objectivity. *Standpoint epistemologies* stress the social situatedness of all perspectives as well as scientific standards. *Feminist postmodernism* doubts the epistemic authority of science in general.

7 It is important to note that a number of critics see an even greater role for values in the determination of mental disorders. For example, Lilienfeld and Marino (1995) and Murphy and Woolfolk (2000) argue that Wakefield's analysis is based on a too simplistic account of evolution and therefore fails to establish a value-free notion of mental dysfunction.

8 Assessing the validity of diagnostic categories is a complex issue; as we cannot directly read such categories of nature (or patients), validity is rather evaluated via its different indicators, or "validators." The "Guidelines for Making Changes to DSM-V" list a number of validators that form different groups. The group of antecedent validators includes factors such as "familial aggregation" or the presence of "environmental risk factors"; predictive validators include, for example, "course of illness" or "response to treatment" (Kendler et al. 2009). The evidential import of these different validators may have to be weighed, as they do not always line up; for example, one way of classifying schizophrenia may fare better in terms of familial aggregation, while another proposal is better in predicting treatment response (Kendler 1990).

References

American Psychiatric Association (APA) (2013), *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Arlington, VA: American Psychiatric Publishing.

Andersen, H. and Wagenknecht, S. (2013), Epistemic Dependence in Interdisciplinary Groups, *Synthese*, 190:1881–1898.

Anderson, E. (2012), Epistemic Justice as a Virtue of Social Institutions, *Social Epistemology*, 26:163–173.

Anderson, E. (2015), Feminist Epistemology and Philosophy of Science. In *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta, ed. URL= https://plato.stanford.edu/archives/fall2015/entries/feminism-epistemology/ Retrieved January 30, 2017.

Biddle, J. and Leuschner, A. (2015), Climate Skepticism and the Manufacture of Doubt: Can Dissent in Science Be Epistemically Detrimental?, *European Journal for Philosophy of Science*, 5:261–278.

Bird, A. (2014), When Is There a Group That Knows? Distributed Cognition, Scientific Knowledge, and the Social Epistemic Subject, pp. 42–63. In *Essays in Collective Epistemology*. J. Lackey, ed. New York: Oxford University Press.

Boorse, C. (1977), Health as a Theoretical Concept, *Philosophy of Science*, 44:542–573.

Brown, J. R. (2001), *Who Rules in Science? An Opiniated Guide to the Wars*, Cambridge: Harvard University Press.

Brown, J. R. (2008), The Community of Science®, pp. 189–216. In *The Challenge of the Social and the Pressure of Practice: Science and Values Revisited*. M. Carrier, D. Howard, and J. Kourany, eds. Pittsburgh, PA: University of Pittsburgh Press.

Bueter, A. (2015), The Irreducibility of Value-Freedom to Theory Assessment, *Studies in History and Philosophy of Science*, 49:18–26.

Bueter, A. (2018), Public Epistemic Trustworthiness and the Integration of Patients in Psychiatric Classification, *Synthese*, online first: https://doi.org/10.1007/s11229-018-01913-z.

Burge, T. (1993), Content Preservation, *Philosophical Review*, 102:457–488.

Carel, H. and Kidd, I. J. (2014), Epistemic Injustice in Healthcare: A Philosophical Analysis, *Medicine, Health Care, and Philosophy*, 17:529–540.

Coady, C. A. J. (1992), *Testimony*, Oxford: Oxford University Press.

Conrad, P. and Potter, D. (2000), From Hyperactive Children to ADHD Adults: Observations on the Expansion of Medical Categories, *Social Problems*, 47:559–582.

Cooper, R. (2002), Disease, *Studies in History and Philosophy of Biological and Biomedical Sciences*, 33:263–282.

Cosgrove, L. and Krimsky, S. (2012), A Comparison of DSM-IV and DSM-5 Panel Members' Financial Associations with Industry: A Pernicious Problem Persists, *PLoS Medicine*, 9:3.

Cosgrove, L., Krimsky, S., Vijayaraghavan, M., and Schneider, L. (2006), Financial Ties between DSM-IV Panel Members and the Pharmaceutical Industry, *Psychotherapy and Psychosomatics*, 75:154–160.

Crichton, P., Carel, H., and Kidd, I. J. (2017), Epistemic Injustice in Psychiatry, *BJPsych Bulletin*, 41(2):65–70.

Douglas, H. (2000), Inductive Risk and Values in Science, *Philosophy of Science*, 67:559–579.

Duhem, P. (1906), *The Aim and Structure of Physical Theory*, 2nd edition. Princeton, NJ: Princeton University Press.

Dupré, J. (2007), Fact and Value, pp. 27–41. In *Value-Free Science? Ideals and Illusions*, J. Dupré, H. Kincaid, and A. Wylie, eds. Oxford: Oxford University Press.

Elliot, K. C. and MacKaughan, D. J. (2009), How Values in Discovery and Pursuit Alter Theory Appraisal, *Philosophy of Science*, 76 (Proceedings):598–611.

Fricker, E. (1994), Against Gullibility, pp. 125–161. In *Knowing from Words*. B. K. Matilal, and A. Chakrabarti, eds. Dordrecht: Kluwer Academic Publishers.

Fricker, M. (2007), *Epistemic Injustice. Power and the Ethics of Knowing*, Oxford: Oxford University Press.

Gilbert, M. (1989), *On Social Facts*, New York: Routledge.

Goldberg, S. (2010), *Relying on Others: An Essay in Epistemology*, Oxford: Oxford University Press.

Goldman, A. I. (1999), *Knowledge in a Social World*, Oxford: Clarendon Press.

Goldman, A. I. (2001), Experts: Which Ones Should You Trust?, *Philosophy and Phenomenological Research*, 63:85–110.

Goldman, A. I. and Blanchard, T. (2016), Social Epistemology. In *The Stanford Encyclopedia of Philosophy*. E. N. Zalta, ed. URL= https://plato.stanford.edu/archives/win2016/entries/epistemology-social/ Retrieved January 30, 2017.

Grasswick, H. (2016), Feminist Social Epistemology. In *The Stanford Encyclopedia of Philosophy*. E. N. Zalta, ed. URL= https://plato.stanford.edu/archives/win2016/entries/feminist-social-epistemology/ Retrieved January 30, 2017.

Harding, S. (1986), *The Science Question in Feminism*, Ithaca, NY: Cornell University Press.

Hardwig J. (1985), Epistemic Dependence, *Journal of Philosophy*, 82:335–349.

Hawthorne, S. C. C. (2010), Institutionalized Intolerance of ADHD: Sources and Consequences, *Hypatia*, 25:504–526.

Hess, D. J. (1997), *Science Studies: An Advanced Introduction*, New York: New York University Press.

Hume, D. (1748), *An Enquiry Concerning Human Understanding*. T. Beauchamp, ed. Oxford: Oxford University Press.

Jukola, S. (2015), Meta-Analysis, Ideals of Objectivity, and the Reliability of Medical Knowledge, *Science & Technology Studies*, 28:101–120.

Kendler, K. S. (1990), Toward a Scientific Psychiatric Nosology: Strengths and Limitations, *Archives of General Psychiatry*, 47:969–973.

Kendler, K., Kupfer, D., Narrow, W., Phillips, K., and Fawcett, J. (2009), Guidelines for Making Changes to DSM-V. URL= www.dsm5.org/ProgressReports/Documents/Guidelines-for-Making-Changes-to-DSM_1.pdf Retrieved September 3, 2014.

Kitcher, P. (1993), *The Advancement of Science: Science without Legend, Objectivity without Illusions*, New York: Oxford University Press.

Kitcher, P. (2001), *Science, Truth, Democracy*, Oxford: Oxford University Press.

Kitcher, P. and Reiss, J. (2010), Biomedical Research, Neglected Diseases, and Well-Ordered Science, *Theoria: An International Journal for Theory, History, and Foundations of Science*, 24:263–282.

Koertge, N. (2000), Science, Values, and the Value of Science, *Philosophy of Science*, 67 (Proceedings):45–57.

Kuhn, T. S. (1962), *The Structure of Scientific Revolutions*, Chicago, IL: University of Chicago Press.

Kusch, M. (2004), *Knowledge by Agreement: The Programme of Communitarian Epistemology*, Oxford: Oxford University Press.

Lackey, J. (2008), *Learning from Words: Testimony as a Source of Knowledge*, Oxford: Oxford University Press.

Lane, C. (2007), *Shyness: How Normal Behavior Became a Sickness*, New Haven, CT: Yale University Press.

Laudan, L. (1990), Demystifying Underdetermination, pp. 267–297. In *Scientific Theories*. C. W. Savage, ed. Minneapolis: University of Minnesota Press.

Lilienfeld, S. O. and Marino, L. (1995), Mental Disorder as a Roschian Concept: A Critique of Wakefield's "Harmful Dysfunction" Analysis, *Journal of Abnormal Psychology*, 104:411–420.

List, C. and Pettit, P. (2002), Aggregating Sets of Judgments: An Impossibility Result, *Synthese*, 140:207–235.

Longino, H. E. (1990), *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*, Princeton, NJ: Princeton University Press.

Longino, H. E. (1996), Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy, pp. 39–58. In *Feminism, Science, and the Philosophy of Science*. L. H. Nelson and J. Nelson, eds. Dordrecht: Kluwer Academic Publishers.

Longino, H. E. (2002), *The Fate of Knowledge*, Princeton, NJ: Princeton University Press.

Lundh, A., Sismondo, S., Lexchin, J., Busuioc, O. A., and Bero, L. (2015), Industry Sponsorship and Research Outcome, *Cochrane Database of Systematic Reviews*, 12.

Murphy, D. and Woolfolk, R. (2000), The Harmful Dysfunction Analysis of Mental Disorder, *Philosophy, Psychiatry, and Psychology*, 7:241–251.

Musschenga, A., Van Der Steen, W., and Ho, V. (2010), The Business of Drug Research: A Mixed Blessing, pp. 110–131. In *The Commodification of Academic Science*. H. Radder, ed. Pittsburgh: Pittsburgh University Press.

Nelson, L. H. (1990), *Who Knows: From Quine to a Feminist Empiricism*, Philadelphia, PA: Temple University Press.

Nordenfelt, L. (1987), *On the Nature of Health: An Action-Theoretic Approach*, Dordrecht: Reidel.

Ohkrulik, K. (1994), Gender and the Biological Sciences, pp. 192–208. In *Philosophy of Science: The Central Issues*. M. Curd and J. A. Cover, eds. New York: Norton 1998.

Oreskes, N. and Conway, E. M. (2010), *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*, London: Bloomsbury.

Porter, D. (2013), Establishing Normative Validity for Scientific Psychiatric Nosology: The Significance of Integrating Patient Perspectives, pp. 63–74. In *Making the DSM-5: Concepts and Controversies*, Springer: New York.

Potter, E. (1996), *Underdetermination Undeterred*, pp. 121–138. In *Feminism, Science, and the Philosophy of Science*. L. H. Nelson and J. Nelson, eds. Dordrecht: Kluwer Academic.

Proctor, R. N. and Schiebinger, L., eds. (2006), *Agnotology: The Making and Unmaking of Ignorance*, Stanford, CA: Stanford University Press.

Quine, W. v. O. (1951), Two Dogmas of Empiricism, pp. 20–46. In *From a Logical Point of View*, 2nd edition. Cambridge: Harvard University Press.

Regier, D. A., Kuhl, E. A., Kupfer, D. J., and McNulty, J. P. (2010), Patient Involvement in the Development of DSM-V, *Psychiatry*, 73:308–310.

Reichenbach, H. (1938), *Experience and Prediction: An Analysis of the Foundations and Structure of Knowledge*, Chicago, IL: Chicago University Press.

Reid, T. (1764), *An Inquiry into the Human Mind*. D. R. Brooks ed. 1997, Edinburgh: Edinburgh University Press.

Ross, P. A. (2002), Values and Objectivity in Psychiatric Nosology, pp. 45–55. In *Descriptions and Prescriptions: Values, Mental Disorders, and the DSMs*. J. Z. Sadler, ed. Baltimore, MD: Johns Hopkins University Press.

Sadler, J. Z. (2005), *Values and Psychiatric Diagnosis*, Oxford: Oxford University Press.

Sadler, J. Z. Fulford, B. (2004), Should Patients and Their Families Contribute to the DSM-5 Process?, *Psychiatric Services*, 55:133–138.

Schickore, J. and Steinle, F., eds. (2006), *Revisiting Discovery and Justification: Historical and Philosophical Perspectives on the Context Distinction*, Dordrecht: Springer.

Selinger, E. and Crease, R. P., eds. (2006), *The Philosophy of Expertise*, New York: Columbia University Press.

Sismondo, S and Doucet, M. (2010), Publication Ethics and the Ghost Management of Medical Publications, *Bioethics*, 24:273–283.

Sisti, D., and Johnson, R. (2015), Revision and Representation: The Controversial Case of DSM-5, *Public Affairs Quarterly*, 29:76–108.

Sisti, D., Young, M., and Caplan, A. (2013), Defining Mental Illnesses: Can Values and Objectivity Get Along?, *BMC Psychiatry*, 13:346.

Solomon, M. (2001), *Social Empiricism*, Cambridge, MA: MIT Press.

Solomon, M. (2015a), *Making Medical Knowledge*, Oxford: Oxford University Press.

Solomon, M. (2015b), Expert Disagreement and Medical Authority, pp. 60–72. In *Philosophical Issues in Psychiatry III: The Nature and Sources of Historical Change*. K. Kendler and J. Parnas, eds. Oxford: Oxford University Press.

Solomon, M. and Kendler, K. S. (2016), Expert Consensus v. Evidence-Based Approaches in the Revision of the DSM, *Psychological Medicine*, 46:2255–2262.

Stein, D. J. and Phillips, K. A. (2013), Patient Advocacy and DSM-5, *BMC Medicine*, 11:133.

Tekin, Ş. (2014), Psychiatric Taxonomy: At the Crossroads of Science and Ethics, *Journal of Medical Ethics*, 40:513–514.

Wakefield, J. (1992), The Concept of Mental Disorder: On the Boundary between Biological Facts and Social Values, *American Psychologist*, 47:373–388.

Wilholt, T. (2009), Bias and Values in Scientific Research, *Studies in History and Philosophy of Science*, 40:92–101.

Study Questions

1. *Are mental disorders real?*
2. *Are mental disorders natural kinds?*
3. *What is epistemic justice in the context of psychiatry?*
4. *How can we know that some people have mental disorders?*
5. *What are the standards for objectivity in psychiatric science?*

Part Eight

Philosophy of Medicine and Psychiatry

25

The Functions of Diagnoses in Medicine and Psychiatry

Hane Htut Maung

Introduction

Diagnoses are central to the practice of medicine, where they serve a variety of functions for clinicians, patients, and society. They aid communication, explain symptoms, inform predictions, guide therapeutic interventions, legitimize sickness, and authorize access to resources. Insofar as psychiatry is a discipline whose practice is shaped by medical conventions, its diagnoses are sometimes presented as if they serve the same sorts of function as diagnoses in bodily medicine. However, there are philosophical problems that cast doubt on whether the functions of psychiatric diagnoses can legitimately be considered to be equivalent to those of medical diagnoses. The aim of this chapter is to explicate some of these problems, particularly conceptual and epistemological problems pertaining to the roles of diagnoses as explanations. I begin with an overview of the various functions that medical diagnoses normally serve and suggest that many of these functions receive justificatory support from the explanatory roles of the diagnoses. I then present issues regarding the epistemic functions of psychiatric diagnoses and how these issues have featured in the arguments of prominent critics of psychiatry.

Diagnoses in medicine

The purpose of this section is twofold. In the first section, I present a general overview of some of the various functions that diagnoses often serve in clinical practice. Then, I look more specifically at the explanatory function and argue

that it is important for providing justificatory support for the other functions of the diagnosis. I focus here mostly on diagnoses in bodily medicine, which will provide a point of comparison for my discussion of psychiatric diagnoses later in the section "The trouble with psychiatry."

The functions of medical diagnoses

(i) *Hypothesis*. The clinical consultation between patient and clinician usually begins with the clinician taking a history from the patient to elicit his or her symptoms and other relevant information, examining the patient to elicit any signs, and reviewing any available investigation results (Stanley and Campos 2013). In practice, one or more of these steps may be omitted, depending on the particular scenario. For instance, in an emergency scenario involving loss of awareness, the patient is unable to provide a history, and the clinical team has to rely on examination signs and investigation results to make a diagnosis. Conversely, in general practice, many diagnoses are informed by the symptoms and signs, without laboratory or radiological investigations being requested. Nonetheless, these minor differences aside, the diagnostic process normally begins with the gathering of a flexible combination of symptoms, signs, and investigation results, henceforth referred to as patient data.

After the patient data is gathered and consolidated, a diagnosis is inferred from the patient data. Further investigations may then be undertaken to acquire evidence that could support or undermine the diagnosis. Usually, several possible diagnoses are initially stipulated and further assessment is undertaken to help select the correct diagnosis from the list of possibilities, a practice known as differential diagnosis (Longmore et al. 2014: 13). The diagnosis, then, functions as a testable hypothesis about the patient's condition that is informed by the patient data. Indeed, several authors have commented on the similarity between the diagnostic process in medicine and hypothesis formation in science, and consider medical diagnoses to be akin to scientific hypotheses (Aliseda and Leonides 2013; Stanley and Campos 2013; Willis, Beebee and Lasserson 2013). For example, it has been suggested that although its conditions are less controlled than those in a laboratory, the diagnostic process is "an example of science in action" (Willis, Beebee and Lasserson 2013: 501).

(ii) *Explanation*. We have just seen that the diagnostic process is akin to scientific hypothesis formation, whereby the diagnosis is inferred from a set of patient data, consisting of symptoms, signs, and investigation results. In turn, it is often the case that the diagnosis explains the patient data. The idea that

diagnoses in medicine often and ideally function as explanations of patients' symptoms is generally accepted in the philosophy of medicine:

> Discomfort makes the patient think that something is wrong with him, and a why-question arises in his mind. … He complains to the physician of these symptoms. … All these clinical manifestations (symptoms, signs and laboratory data) require an explanation from the physician, and finally a diagnosis is reached. (Qiu 1989: 199)

> When a patient goes to a physician with a set of complaints and symptoms, the physician's first task is to make a diagnosis of a disease that explains the symptoms. (Thagard 1999: 20)

> Once formulated, however, a diagnosis can be synthetically described, from a statistical viewpoint, as a relation between a set of findings (signs, symptoms, laboratory test results) and a certain pathological condition attributed to the patient. What kind of relation? According to a common opinion among experts in computational models, medical diagnoses express *explanatory* relations. (Benzi 2011: 365)

Moreover, it is generally accepted that the sort of explanation is normally a causal explanation, whereby the diagnosis explains the patient data by indicating its cause:

> To solve a clinical diagnostic problem means first to recognize a malfunction and then to set about tracing or identifying its causes. The diagnosis is thus an explanation of disordered function, where possible a causal explanation. (Schwartz and Elstein 2008: 224)

> It is uncontroversial in the medical literature that the ideal diagnosis is a biomedical causal explanation. … Such a diagnosis posits a physiological cause for a set of physical signs and symptoms. (Cournoyea and Kennedy 2014: 928–929)

And so, there is often a bidirectional epistemic relation between the diagnosis and the patient data. The diagnosis is inferred from the patient data and the patient data is explained by the diagnosis. Later in the subsection "Why explanation is important." I suggest that this explanatory function of a diagnosis is important, because it provides justificatory support for its other functions, such as prediction and intervention.

(iii) *Prediction*. In addition to post hoc explanation of patient data, a diagnosis serves a predictive function. The clinician is very often able to make reliable predictions about the likely future outcome for a patient based on the diagnosis. These include predictions about the prognosis, which consists of the clinical course and likelihood of survival, predictions about potential complications,

and predictions about responses to treatments. For instance, the diagnosis of acute appendicitis informs the clinician that the patient's condition is likely to deteriorate rapidly without treatment, that a potential complication is appendiceal perforation, and that a good recovery is likely following surgical removal of the appendix. Therefore, a diagnosis serves the epistemic function of supporting inferences about the future.

(iv) *Intervention*. It is uncontroversial that an important function of a diagnosis is to guide intervention. Indeed, theorists have proposed that the value of the diagnosis must be considered relative to the therapeutic goals of medicine. Caroline Whitbeck argues that the diagnosis is "aimed at obtaining the best medical outcome for the patient" (Whitbeck 1981: 326), while Annemarie Jutel (2011: 21) notes that the diagnostic process is very often motivated by the goal to ascertain the correct treatment. Therefore, a diagnosis not only has epistemic value, but also instrumental utility in guiding treatment, which makes it a key component of practical reasoning in medicine.

As alluded to above, this interventional function of a diagnosis is supported by its predictive function. A diagnosis can inform predictions about likely responses to treatments, and so can guide therapeutic decision making. For example, a clinician can predict from the diagnosis of acute appendicitis that the patient is likely to make a good recovery following an operation to remove the appendix, thus supporting the decision to intervene therapeutically with an operation to remove the appendix.

(v) *Denotation*. As well as the abovementioned epistemic functions, the hypothesized diagnosis serves a linguistic function as a denotative label for the condition with which the patient is presenting. It comprises a term that is understood to refer to a state of affairs in the patient, such as "acute appendicitis" referring to the rapid onset of inflammation of the appendix. Furthermore, as noted by physician and psychoanalyst Michael Balint (1964: 25), such a diagnostic term provides a useful shorthand description that organizes a variety of disparate clinical features into a unified phenomenon. This is important because it facilitates the communicative exchanges of clinicians. Hence, diagnostic terms constitute part of a common language with which clinicians can reliably and concisely convey clinical information to each other.

(vi) *Classification*. Denotation is closely related to classification. Designating a condition with a specific term implies conceptually distinguishing it from other conditions. The diagnosis of acute appendicitis specifically denotes inflammation of the appendix, which is taken to be conceptually and empirically distinct from, for example, inflammation of the gallbladder or obstruction of the

bowel. Moreover, the diagnostic term is not merely taken to denote an individual instance of the condition, but represents a generalized category. Hence, the condition denoted by the diagnosis is often considered to be a repeatable type, of which individual cases are tokens (Sadler 2005: 419–420).

Diagnostic terms, then, demarcate and classify diseases into clinically significant categories. Hence, as Annemarie Jutel claims, the diagnosis is "one of medicine's most powerful classification tools" (Jutel 2011: 15). This is reflected by the profound and pervasive influences of formal diagnostic classification systems on public policy, health insurance, and pharmaceutical research (Cooper 2005: 1). Two of the leading formal diagnostic classification systems in current usage are the World Health Organization's *International Classification of Diseases* (*ICD*), now in its tenth revision (1992), and, in the field of psychiatry, the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*), now in its fifth edition (2013). However, even outside the official taxonomies of *ICD*-10 and *DSM*-5, the classificatory functions of diagnoses are deeply embedded in everyday clinical practice. For example, the *Oxford Handbook of Clinical Medicine* (Longmore et al. 2014), which is considered an indispensable resource for medical students and physicians, organizes diagnoses into cardiovascular disorders, respiratory disorders, gastrointestinal disorders, endocrine disorders, infectious diseases, malignancies, and so on.

(vii) *Normative*. The functions discussed so far have been largely descriptive, that is, they concern the role of the diagnosis in picking out a biological state of affairs that is assumed to be part of the external world, albeit occurring within the body of the patient. However, the diagnosis also has a normative function. Assigning a diagnosis to a patient does not only pick out a state of affairs, but usually implies the evaluative judgment that this state of affairs is abnormal. More specifically, the diagnosis usually implies that the patient has a medical disorder. According to Jerome Wakefield (1992) it is important that a diagnostic category discerns cases of genuine disorder from non-disordered cases, such as variants of normality, and from other sorts of problem, such as social and moral problems. This normative function of a diagnosis is often used to offer vindication for some of the other functions of the diagnosis. With respect to the interventional function, for example, the normative judgment regarding whether the patient's condition is a disorder informs the decision about whether medical intervention is appropriate at all.

(viii) *Semiotic*. So far, I have discussed functions of a diagnosis that are useful for the clinician. However, a diagnosis can also serve a function for the patient. More specifically, it functions as a "semiotic mediator," or a meaningful label

which the patient can use to understand and act upon his or her condition (Brinkmann 2014). For example, a diagnosis could be taken by the patient as legitimizing his or her illness, thus validating his or her personal experience of being unwell as something that deserves to be taken seriously. This draws on the abovementioned normative function of the diagnosis. More broadly, Carl Elliott (1999) proposes that a diagnosis can influence the narrative by which one interprets one's life and shapes one's future. When the effect of the diagnosis on the one's life narrative is significant, such as with a chronic, untreatable, or potentially fatal condition, it can profoundly reorganize one's sense of personal identity and attitude toward what is valuable in life (Kleinman 1988; Tekin 2011; 2014).

Often, the semiotic function served by diagnosis can be helpful for the patient as it can enable the patient to plan his or her life accordingly. This can be the case even if the diagnosis is of a serious condition. For instance, polycystic kidney disease is an autosomal-dominant inherited disease associated with progressive renal failure and a significantly increased risk of subarachnoid hemorrhage. A diagnosis of polycystic kidney disease could enable the patient to take measures to control his or her blood pressure, attend regular neuroimaging scans to screen for cerebral aneurysms, consider the possible need for dialysis in the future, and make an informed decision about family planning in light of it being possible that his or her children could inherit the disease.

(ix) *Social*. Finally, a diagnosis functions as a social performative that influences attitudes and behaviors at individual, institutional, and cultural levels. At the individual level, John Sadler (2005: 421–422) notes that the diagnosis endows the clinician with certain privileges. These might include initiating pharmacological treatment, surgical intervention, psychological therapy, and potentially invasive testing. Hence, the social function of the diagnosis is closely related to its interventional and normative functions. At an institutional level, a diagnosis entitles the patient to therapeutic, supportive, and financial resources to which he or she had not previously been entitled. For example, the diagnosis of myocardial infarction entitles the patient to a hospital bed, nursing care, laboratory and radiological investigations, medical and surgical interventions, rehabilitation, and outpatient follow-up after discharge into the community. When the illness is more chronic and disabling, a diagnosis can also authorize the patient's access to further supportive and financial resources, including attendance to support groups, carer input, disability benefits, and supported accommodation. At the level of culture, a diagnosis legitimizes sickness and sanctions certain kinds of behavior (Jutel 2011: 7). The sociologist Talcott

Parsons (1951: 436–437) proposes that the patient is thrust into a "sick role," which bestows on him or her certain rights and duties. The patient's rights are to not be considered responsible for his or her illness and to be exempt from some of his or her normal obligations, while his or her duties are to try to get well and to seek appropriate medical care.

Why explanation is important

The various functions discussed in the subsection "The functions of medical diagnoses" make the diagnosis a valuable epistemic resource in clinical practice. In this current subsection, I focus my attention specifically on the explanatory function of the diagnosis. I suggest that this explanatory function is important because it provides justificatory support for many of the other functions. My claim is not that the explanatory function is necessary for these other functions, but the more modest proposal that these other functions are strengthened by the explanatory function of the diagnosis.

There is a clear connection between the function of a diagnosis as a hypothesis and its function as an explanation. In general, when we infer hypothesis from a set of data, we want the hypothesis to explain the data. This squares with the idea that the diagnostic process involves abductive reasoning, or inference to the best explanation (Aliseda and Leonides 2013; Stanley and Campos 2013). Moreover, as noted by Peter Lipton (2004), explanatory power is a value that is used to judge the quality of the hypothesis. Hence, explanatory considerations not only motivate and guide the inferential process in diagnostic hypothesis formation, but are appealed to in the evaluation of the hypothesis.

The explanatory function of a diagnosis is also closely connected to its denotative and classificatory functions. Where a diagnosis serves as an explanation of patient data, it does so partly by denoting a kind of causal structure that is instantiated by the actual patient. For example, the diagnosis of acute appendicitis explains a patient's abdominal pain by denoting a distinctive pathological type, in this case acute inflammation of the appendix, which is causing the abdominal pain. Conversely, causal explanatory considerations partly justify why some conjunctions of clinical phenomena, but not others, are made into diagnostic categories and assigned diagnostic terms. According to Neil Williams (2011), it is often the case that when seemingly disparate clinical phenomena are clustered together and characterized as a distinctive category, it is because scientists and medical professionals the clinical phenomena to be connected by a unifying causal explanation. Indeed, a diagnostic category can

be discarded and replaced by more precise categories if it turns to be too causally heterogeneous to serve as a satisfactory causal explanation, such as dropsy being discarded and replaced by the more precise categories of congestive cardiac failure, cirrhosis of the liver, and nephrotic syndrome. Hence, the causal explanatory value of a diagnostic category influences our judgments about the validity of the classification. As we shall see in the section "The trouble with psychiatry," this is apparent in some of the recent philosophical critiques of diagnostic classification in psychiatry (Murphy 2006; Poland 2014).

The predictive and interventional functions of a diagnosis are supported by its explanatory function, in particular its causal explanatory function. To be clear, this is not to say that causal explanation is necessary for successful prediction or intervention. As noted by Jennifer Radden (2003: 46), a diagnostic category that is defined by a cluster of symptoms without allusion to an underlying cause can still permit probabilistic predictions. We may not know what causes this cluster of symptoms, but we could nonetheless make inductive inferences about similar cases based on enumerative induction, which can then inform an evidenced-based treatment guideline. Nonetheless, Radden also argues that a diagnosis that is explanatory is superior to one that is merely descriptive, because it opens up possibilities for further hypotheses and targeted interventions. Explaining why a patient has a particular cluster of symptoms provides understanding of the underlying causal structure, which can signal potential targets for therapeutic interventions, inform decisions regarding treatment approaches, and allow us to make predictive inferences that go beyond enumerative induction.

The explanatory function of a diagnosis is often considered to justify its normative and social functions. As noted by Annemarie Jutel, a diagnosis "explains certain kinds of deviance in terms of disease rather than of moral failing" (Jutel 2011: 229). This is then regarded as a reason to excuse the patient from certain responsibilities and grant him or her certain rights according to the "sick role" (Parsons 1951). For instance, a child diagnosed with influenza may be temporarily granted absence from school, because his or her failure to concentrate is explained as being due to an unpleasant and unfortunate medical problem, rather than deliberate school refusal. Furthermore, by indicating that a patient's symptoms are caused by a particular kind of condition, the explanatory function of a diagnosis can support the mobilization of therapeutic and supportive resources of the sorts deemed by medical professionals and policymakers to be beneficial for the particular kind of condition.

Finally, the function of the diagnosis as a semiotic mediator draws on its explanatory function. Part of why a diagnosis serves as a meaningful sign for the

patient is because it is taken to provide an explanation of why he or she has been suffering from his or her symptoms. Kirmayer et al. (2004: 664) suggest that "explanations may offer some reassurance and consolidation, promote coping and resilience, and allow the person to plan realistically for the future." Similarly, Winston Chiong (2004: 129) writes that a diagnosis "can also be an explanation for patients who have had symptoms but do not know their cause," which "may seem to resolve the mystery, such that even patients with intractable, chronic diseases may feel relief when diagnosed." And so, the explanatory function of a diagnosis does not only have epistemic significance and instrumental utility for the clinician, but also has intrinsic value for the patient.

The trouble with psychiatry

Having laid out the roles of medical diagnoses as causal explanations of symptoms, I now turn to issues concerning the functions of psychiatric diagnoses. In the first subsection, I show how diagnoses in psychiatry are sometimes portrayed in clinical discourse as if they serve as causal explanations in ways analogous to diagnoses in bodily medicine. In the next two subsections, I present two problems, respectively derived from two arguments by Thomas Szasz (1960), which raise doubts about the legitimacy of invoking psychiatric diagnoses as explanations. I respectively call these the conceptual problem and the ontological problem. Finally, I look at how these problems have featured in more recent normative critiques of psychiatry by mental health practitioners and philosophers of psychiatry. I do not, in this chapter, provide solutions to these problems, as my intention here is merely to explicate them, but I have offered philosophical solutions to the ontological and conceptual problems elsewhere (Maung 2016a; b).

The uses of diagnoses in psychiatric discourse

I showed in the previous two sections that medical diagnoses often explain patients' symptoms by indicating the underlying conditions responsible for causing the symptoms. In clinical discourse, psychiatric diagnoses are sometimes portrayed as if they also serve such causal explanatory roles. This is perhaps unsurprising, given the historical and cultural underpinnings of psychiatry as a medical discipline. As noted by Jeffrey Poland (2014), psychiatry is a discipline

whose practice occurs in a context shaped by medical traditions. Hence, it is understandable that its practitioners apply to it the methods and conventions of other medical disciplines.

Such portrayals of psychiatric diagnoses as causes or explanations of symptoms can be found in clinical textbooks, as shown by the following passages:

> The diagnosis of antisocial personality disorder is not warranted if the symptoms can be *explained by* schizophrenia, mania or mental retardation. (Sethi 2008: 109, italics added)

> As an example, think of the differential diagnosis of a patient with episodes of anxiety and breathlessness. These symptoms are often *caused by* panic disorder. (Stevens and Rodin 2010: 74, italics added)

> Depression and anxiety *cause* tiredness as do some somatization disorders. … Anaemia, liver failure, coeliac disease, cancer, Parkinson's, alcohol overdose and rare disorders such as myasthenia gravis and motor neurone disease can also cause tiredness. (Wright, Dave and Dogra 2010: 152, italics added)

Similarly, *A Guide to Psychiatric Examination* lists schizophrenia, mania, and depression alongside dementia and medical conditions as "common causes of psychoses" (Aquilina and Warner 2004: 79). These passages suggest that psychiatrists are encouraged to think about diagnoses in psychiatry as being analogous to diagnoses in bodily medicine.

Such causal claims can also be found in health information resources about psychiatric disorders that are targeted at the general public. Consider the following passages about schizophrenia and major depressive disorder from Patient.info and NHS Choices:

> Schizophrenia is a serious mental health condition that *causes* disordered ideas, beliefs and experiences. (Patient.info 2013, italics added)

> Depression affects people in different ways and can *cause* a wide variety of symptoms. They range from lasting feelings of sadness and hopelessness, to losing interest in the things you used to enjoy and feeling very tearful. (NHS Choices 2014: italics added)

Such portrayals of psychiatric diagnoses are significant, because they can influence how patients perceive and respond to their conditions. In a qualitative study of adults diagnosed with attention-deficit hyperactivity disorder, Svend Brinkmann (2014) notes that the participants commonly mediate understanding of their problematic behaviors by invoking their diagnoses as explanations of these behaviors. Hence, there is some evidence that patients

think of psychiatric diagnoses as if they refer to underlying conditions that cause symptoms.

The conceptual problem

The passages quoted in the previous subsection suggest that psychiatric diagnoses are sometimes communicated to clinicians, patients, and the public as if they are causal explanations of patients' symptoms, much like the diagnoses in other medical specialties. However, there are worries about psychiatric diagnoses that raise doubts whether they actually do explain symptoms. One such worry concerns the way that psychiatric diagnoses are defined. I call this the conceptual problem. According to the most recent editions of *DSM*, psychiatric diagnoses are formally defined in terms of their symptoms. Consider the following excerpts from *DSM*-5:

> The *essential feature* of a major depressive episode is a period of at least 2 weeks during which there is either depressed mood or the loss of interest or pleasure in nearly all activities. (American Psychiatric Association 2013: 163, italics added)

> Panic disorder *refers to* recurrent unexpected panic attacks. A panic attack is an abrupt surge of intense fear or intense discomfort that reaches a peak within minutes. (American Psychiatric Association 2013: 209, italics added)

> The *essential feature* of generalized anxiety disorder is excessive anxiety and worry (apprehensive expectation) about a number of events or activities. (American Psychiatric Association 2013: 222, italics added)

These descriptive definitions in *DSM*-5 suggest that psychiatric diagnoses are constituted by their symptoms. It is fairly uncontroversial that causes are distinct from their effects. That is to say, something cannot be its own cause. Therefore, if psychiatric diagnoses are mere labels for clusters of symptoms, then they cannot refer to the causes of these symptoms.

While the *DSM* formalized the descriptive approach to defining psychiatric diagnoses, the worry that psychiatric diagnoses merely have definitional connections with their respective symptoms had been present even before the introduction of the fully descriptive nosology in *DSM*-III. In "The Myth of Mental Illness" (1960), the psychiatrist and leading figure of the antipsychiatry movement Thomas Szasz presents two arguments against the validity of the concept of mental illness. The first argument is that mental illness diagnoses are mere shorthand labels for certain kinds of behavior, and so cannot also be invoked to refer to causes of these behaviors:

> This is obviously fallacious reasoning, for it makes the abstraction "mental illness" into a cause, even though this abstraction was created in the first place to serve only as a shorthand expression for certain types of human behaviour. (Szasz 1960: 114)

The second argument is that mental illnesses are not genuine disorders because, unlike bodily illnesses, they are not characterized by pathophysiological lesions, but by deviations from social and moral norms. I turn to the second argument in my discussion of the ontological problem in the subsection "The ontological problem," but it is the first argument that is at the core of the conceptual problem.

Tim Thornton (2007: 16) offers an interpretation of Szasz's first argument in terms of necessity and contingency, with reference to David Hume's ([1748] 2000) analysis of causation. According to Hume, causal connections are contingent. We perceive causes and effects as distinct events, but do not perceive any necessary connection between them. Even if the causal chain is broken down further, we only perceive a finer succession of distinct causes and effects, but not any glue between them. Hence, one can conceive one event occurring without the other. For example, while it may be the case that a particular patient's abdominal pain is caused by acute appendicitis, it is conceivable that acute appendicitis could occur without abdominal pain or that abdominal pain could occur without acute appendicitis. However, if a psychiatric diagnosis is defined by its symptoms, then the connection between the diagnosis and the symptoms is not contingent, but necessary. One cannot have panic disorder without having panic attacks, or generalized anxiety disorder without having excessive anxiety. Since contingency is an essential feature of causal connections, it follows that the connection between a psychiatric diagnosis and its symptoms is not causal.

Szasz's argument can also be hinged at the level of language, and examined in terms of analyticity and syntheticity. According to Immanuel Kant ([1781] 1998), an analytic proposition is true in virtue of its meaning, as its predicate concept is contained in its subject concept. A classic example is the proposition "all bachelors are unmarried." This proposition is analytically true, because the concept "unmarried" is contained in the concept "bachelor." By contrast, a synthetic proposition can only be true in virtue of its relation to the state of affairs in the world, because its predicate concept is not contained in its subject concept. For example, the proposition "all bachelors are unhappy," is synthetic, because the concept "unhappy" is not contained in the concept "bachelor." Applied to diagnoses, the proposition "this patient with acute appendicitis has abdominal pain," is synthetic, because the concept "abdominal pain" is not contained in the concept "acute appendicitis." However, the proposition "this

patient with panic disorder has recurrent unexpected panic attacks," is analytic, because, according to the *DSM*-5 definition, the concept "recurrent unexpected panic attacks" is contained in the concept "panic disorder." Again, this suggests that the relations between psychiatric diagnoses and their symptoms are not empirical, but definitional.

The above considerations raise serious doubts about whether psychiatric diagnoses can serve the same causal explanatory functions as medical diagnoses. Jennifer Radden (2003) notes that while the purely descriptive approach to defining and classifying psychiatric diagnoses in the most recent editions of the *DSM* does permit probabilistic predictions, it renders its diagnostic categories devoid of explanatory value. Because the connection between a psychiatric diagnosis and its symptoms is definitional rather than causal, such a diagnosis does not explain its symptoms, but merely describes them. The diagnosis of panic disorder does not explain why a patient has recurrent and unexpected panic attacks any more than a man's bachelorhood explains why he is unmarried.

The ontological problem

While the conceptual problem concerns the semantic relations between diagnostic terms and descriptions of symptoms, the ontological problem raises doubts about the causal bases of psychiatric syndromes. This is related to Szasz's second argument, to which I alluded earlier. Szasz argues that mental illnesses are not genuine illnesses, because they are not characterized by distinctive pathophysiological lesions:

> The crux of the matter is that a disease of the brain, analogous to a disease of the skin or bone, is a neurological defect … For example, a defect in a person's visual field may be satisfactorily explained by correlating it with certain definite lesions in the nervous system. On the other hand, a person's belief … cannot be explained by a defect or disease of the nervous system. (Szasz 1960: 113)

It is perhaps fair to say that our scientific understanding of some psychiatric disorders has progressed since the publication of Szasz's paper. However, Szasz's argument still resonates strongly. Empirical research has revealed an array of causes associated with many of the major psychiatric syndromes, but the story has not been one of definite lesions. Rather, it has been one of complexity and heterogeneity at multiple levels of analysis, including the biological, psychological, and social (Bolton 2012; Kendler 2008; Murphy 2006). Hence, it may be that a given diagnostic category in psychiatry does not correspond to a distinctive

causal structure, but is associated with a range of possible causal pathways, each involving complex interactions of diverse factors across different levels.

This causal heterogeneity suggests that a psychiatric diagnosis does not pick out a distinctive structure or process that is citable as a cause in a causal explanation. Rather, it subsumes a range of possible causal structures, each made up of varying combinations of biological, psychological, and social factors. In other words, in different patients with the same diagnosis, the symptoms may be caused by different sorts of process. This raises doubts about the explanatory value of a psychiatric diagnosis, as it suggests that the diagnosis does not specify what is causing the patient's symptoms. Therefore, it falls short of the paradigmatic case in bodily medicine where the diagnosis explains a patient's symptoms by indicating the cause.

Recent criticism of psychiatric diagnosis

The problems explicated above have featured in various forms in recent critiques of psychiatry. In both psychiatry and philosophy, theorists have considered the symptom-based definitions of psychiatric diagnoses and the issue of causal heterogeneity to raise doubts about the validity of psychiatric classification. Kendell and Jablensky write:

> [T]he surface phenomena of psychiatric illness (i.e., the clustering of symptoms, signs, course, and outcome) provide no secure basis for deciding whether a diagnostic class or rubric is valid, in the sense of delineating a specific, necessary, and sufficient biological mechanism. (Kendell and Jablensky 2003: 7)

Similarly, the philosopher Jeffrey Poland (2014) criticizes the epistemic shortcomings of the current psychiatric diagnoses in the *DSM*:

> The *DSM* categories and associated epistemic practices related to information processing, inferential practice, explanatory practice, and clinical understanding, are ineffective and harmfully biased because, given their atheoretical focus on clinical phenomenology, they do not effectively identify and represent important features, problems, contexts, and processes … (i.e., they do not underwrite sound clinical inferences and judgments concerning what is wrong and what is likely to be helpful). (Poland 2014: 48)

Poland's critique suggests that explanatory failure is connected to other shortcomings regarding classification, prediction, and intervention. That is to say, psychiatric diagnoses that do not adequately inform us about the processes underlying patients' problems are unlikely to support reliable inferences or guide

effective treatment decisions. He describes such diagnoses as "free riders" that contribute little over and above descriptions of symptoms (Poland 2014: 34).

In light of these problems, some theorists have suggested that diagnostic classification in psychiatry ought to be revised, so that its diagnostic categories correspond to more stable kinds of causal structure (Murphy 2006; Tsou 2015). According to Dominic Murphy (2006: 323–324), this would bring psychiatry in line with the rest of medicine, where diagnoses correspond to the causal antecedents of symptoms. However, it has also been argued that the high degrees of causal complexity associated with mental disorders pose significant problems for the prospects of such an etiological classification that would be acceptable for practitioners and researchers (Bolton 2012).

There are also concerns regarding the semiotic roles of psychiatric diagnoses. Again, these concerns relate to doubts regarding their explanatory functions. Şerife Tekin (2011, 2014) argues that the symptom-based approach to diagnosis in the *DSM* might contribute to impoverished self-insight because such a descriptive approach ignores the individual factors and circumstances that are relevant to the development of the patient's condition. Moreover, Tekin suggests that the assumption of a biomedical disease model, whereby a given psychiatric category is assumed to reflect a particular kind of biological causal structure, further compounds this impoverished self-insight, because it fails to acknowledge the causal complexity of the disorder and ignores important causal contributory factors, such as psychological features, social context, and interpersonal dynamics.

In addition to the above criticisms offered by philosophers, the epistemic roles of psychiatric diagnoses have also recently been contested by mental health practitioners. Joanna Moncrieff (2010), a psychiatrist and a leading figure of the currently active critical psychiatry movement, criticizes the uses of psychiatric diagnoses in social discourse. Drawing on the work of David Ingleby (1982), Moncrieff argues that psychiatric diagnoses are only allowed to instigate the social responses of mobilizing clinical resources and sanctioning certain behaviors because they are presented by the psychiatric profession as designating diseases that are responsible for the patients' symptoms, much like diagnoses in other medical specialties. However, she suggests that such social responses are not entirely justified, because psychiatric diagnoses do not genuinely designate diseases that explain patients' symptoms.

In light of these concerns, some clinical psychologists have advocated moving away from using categorical diagnoses in psychiatry and have recommended alternative approaches. For example, Richard Bentall (2009) suggests that

assessments of patients should focus on individual complaints, such as paranoia and auditory hallucinations, rather than syndromes, such as schizophrenia and bipolar disorder. Lucy Johnstone (2014) argues that a categorical diagnosis fails to capture important information about the patient's condition and instead recommends the use of a case formulation, which consists of a narrative account that is individualized to the patient's particular case. While Johnstone presents the formulation as an alternative to the diagnosis, the view put forward by the World Psychiatric Association (2003) is that the diagnosis and the formulation have complementary roles. The general idea is that the assessment of the patient ought to include a categorical diagnosis and an individualized formulation, which together comprise a more comprehensive account of the patient's condition.

And so, the issues considered throughout this section highlight some differences between diagnoses in medicine and those in psychiatry. First, diagnoses in medicine are often, though by no means always, defined in terms of the pathologies that produce symptoms, while diagnoses in psychiatry tend to be defined in terms of the clusters of symptoms themselves. Second, the diagnostic categories in medicine tend to correspond to reasonably stable and distinctive causal structures, while there is evidence to suggest that many of those in psychiatry are associated with high degrees of complexity and heterogeneity with respect to their underlying causal processes. These differences are important, because they cast doubt on whether medical diagnoses and psychiatric diagnoses can be considered to serve equivalent epistemic functions. This subsection has looked at some of the concerns raised about the roles of psychiatric diagnoses in explanation, prediction, intervention, classification, and the sanctioning of social responses. In addition to these, I suggest that there are potential ethical implications for psychiatric discourse. As noted in the subsection "The uses of diagnoses in psychiatric discourse," psychiatric diagnoses are often communicated to the public as if they refer to conditions that causally explain symptoms. However, if they do not serve such explanatory functions, then it is likely that patients and the wider public are being misinformed about psychiatric diagnoses. This raises the possibility that patients are misled into believing that their symptoms are being explained, when they are merely being labeled.

Summary

This chapter has explored some of the philosophical problems inherent in the comparisons between the functions of diagnoses in psychiatry and those of

diagnoses in the rest of medicine. In the course of the discussion, I have provided an overview of the various functions that medical diagnoses normally serve, shown how many of these functions receive justificatory support from the roles of the diagnoses as causal explanations of symptoms, presented conceptual and ontological problems concerning the explanatory statuses of psychiatric diagnoses, and looked at how these problems have featured in recent critiques of psychiatry.

References

Aliseda, A. and Leonides, L. (2013), Hypothesis Testing in Adaptive Logics: An Application to Medical Diagnosis, *Logic Journal of the IGPL*, 21: 915–930.

American Psychiatric Association (2013), *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Arlington, VA: American Psychiatric Publishing.

Aquilina, C. and Warner, J. (2004), *A Guide to Psychiatric Examination*, Cheshire: PasTest.

Balint, M. (1964), *The Doctor, His Patient, and the Illness*, 2nd edition. London: Churchill Livingstone.

Bentall, R. P. (2009), *Doctoring the Mind: Why Psychiatric Treatments Fail*, London: Allen Lane.

Benzi, M. (2011), Medical Diagnosis and Actual Causatio', *L&PS—Logic and Philosophy of Science*, 9:365–372.

Bolton, D. (2012), Classification and Causal Mechanisms: A Deflationary Approach to the Classification Problem, pp. 6–11. In *Philosophical Issues in Psychiatry II: Nosology*. K. S. Kendler and J. Parnas, eds. Oxford: Oxford University Press.

Brinkmann, S. (2014), Psychiatric Diagnoses as Semiotic Mediators: The Case of ADHD, *Nordic Psychology*, 66:121–134.

Chiong, W. (2004), Diagnosing and Defining Disease, pp. 128–135. In *Health, Disease, and Illness: Concepts in Medicine*. A. L. Caplan, J. J. McCartney, and D. A. Sisti, eds. Washington, DC: Georgetown University Press.

Cooper, R. (2005), *Classifying Madness: A Philosophical Examination of the Diagnostic and Statistical Manual of Mental Disorders*, Dordrecht: Springer.

Cournoyea, M. and Kennedy, A. G. (2014), Causal Explanatory Pluralism and Medically Unexplained Physical Symptoms, *Journal of Evaluation in Clinical Practice*, 20:928–933.

Elliott, C. (1999), *A Philosophical Disease: Bioethics, Culture and Identity*, New York: Routledge.

Hume, D. (1748/2000), *An Enquiry concerning Human Understanding*, Oxford: Oxford University Press.

Ingleby, D. (1982), The Social Construction of Mental Illness, pp. 123–143. In *The Problem of Medical Knowledge: Examining the Social Construction of Medicine*. P. Wright and A. Treacher, eds. Edinburgh: Edinburgh University Press.

Johnstone, L. (2014), *A Straight Talking Introduction to Psychiatric Diagnosis*, Monmouth: PCCS Books.

Jutel, A. G. (2011), *Putting a Name to It: Diagnosis in Contemporary Society*, Baltimore, MD: Johns Hopkins University Press.

Kant, I. ([1781]1998), *A Critique of Pure Reason*, Cambridge: Cambridge University Press.

Kendell, R. E. and Jablensky, A. (2003), Distinguishing between the Validity and Utility of Psychiatric Diagnoses, *American Journal of Psychiatry*, 160:4–12.

Kendler, K. S. (2008), Explanatory Models for Psychiatric Illness, *American Journal of Psychiatry*, 165:695–702.

Kirmayer, L. J., Groleau, D., Looper, K. J., and Dao, M. D. (2004), Explaining Medically Unexplained Symptoms, *Canadian Journal of Psychiatry*, 49:663–671.

Kleinman, A. (1988), *The Illness Narratives: Suffering, Healing, and the Human Condition*, New York: Basic Books.

Lipton, P. (2004), *Inference to the Best Explanation*, 2nd edition. London: Routledge.

Longmore, M., Wilkinson, I. B., Baldwin, A., and Wallin, E. (2014), *Oxford Handbook of Clinical Medicine*, 9th edition. Oxford: Oxford University Press.

Maung, H. H. (2016a), Diagnosis and Causal Explanation in Psychiatry, *Studies in History and Philosophy of Biological and Biomedical Sciences*, 60:15–24.

Maung, H. H. (2016b), To What Do Psychiatric Diagnoses Refer? A Two-Dimensional Semantic Analysis of Diagnostic Terms, *Studies in History and Philosophy of Biological and Biomedical Sciences*, 55:1–10.

Moncrieff, J. (2010), Psychiatric Diagnosis as a Political Device, *Social Theory and Health*, 8:370–382.

Murphy, D. (2006), *Psychiatry in the Scientific Image*, Cambridge, MA: MIT Press.

NHS Choices (2014), *Clinical Depression*, URL= http://www.nhs.uk/Conditions/Depression/Pages/Introduction.aspx Retrieved April 4, 2016.

Parsons, T. (1951), *The Social System*, Glencoe, IL: The Free Press.

Patient.info. (2013), *Schizophrenia*, Available online: http://patient.info/health/schizophrenia-leaflet Retrieved April 4, 2016.

Poland, J. (2014), Deeply Rooted Sources of Error and Bias in Psychiatric Classification, pp. 29–64. In *Classifying Psychopathology: Mental Kinds and Natural Kinds*. H. Kincaid and J. Sullivan, eds. Cambridge, MA: MIT Press.

Qiu, R. Z. (1989), Models of Explanation and Explanation in Medicine, *International Studies in the Philosophy of Science*, 3:199–212.

Radden, J. (2003), Is This Dame Melancholy? Equating Today's Depression and Past Melancholia, *Philosophy, Psychiatry, and Psychology*, 10:37–52.

Sadler, J. Z. (2005), *Values and Psychiatric Diagnosis*, Oxford: Oxford University Press.

Schwartz, A. and Elstein, A. S. (2008), Clinical Reasoning in Medicine, pp. 223–234. In *Clinical Reasoning in the Health Professions*. J. Higgs, M. A. Jones, S. Loftus, and N. Christensen, eds. 3rd edition. Amsterdam: Elsevier.

Sethi, S. (2008), *Textbook of Psychiatry*, New Delhi: Elsevier.

Stanley, D. E. and Campos, D. G. (2013), The Logic of Medical Diagnosis, *Perspectives in Biology and Medicine*, 56:300–315.

Stevens, L. and Rodin, I. (2010), *Psychiatry: An Illustrated Colour Text*, 2nd edition. London: Churchill Livingstone.

Szasz, T. (1960), The Myth of Mental Illness, *American Psychologist*, 15:113–118.

Tekin, Ş. (2011), Self-Concept through the Diagnostic Looking Glass: Narratives and Mental Disorder, *Philosophical Psychology*, 24:357–380.

Tekin, Ş. (2014), Self-Insight in the Time of Mood Disorders: After the Diagnosis, Beyond the Treatment, *Philosophy, Psychiatry, and Psychology*, 21:139–155.

Thagard, P. (1999), *How Scientists Explain Disease*, Princeton, NJ: Princeton University Press.

Thornton, T. (2007), *Essential Philosophy of Psychiatry*, Oxford: Oxford University Press.

Tsou, J. Y. (2015), DSM-5 and Psychiatry's Second Revolution: Descriptive vs. Theoretical Approaches to Psychiatric Classification, pp. 43–62. In *The DSM-5 in Perspective: Philosophical Reflections on the Psychiatric Babel*. S. Demazeux and P. Singy, eds. Dordrecht: Springer.

Wakefield, J. C. (1992), On the Concept of Mental Disorder: On the Boundary between Biological Facts and Social Values, *American Psychologist*, 47:373–388.

Whitbeck, C. (1981), What Is Diagnosis? Some Critical Reflections, *Metamedicine*, 2:319–329.

Williams, N. E. (2011), Arthritis and Nature's Joints, pp. 199–230. In *Carving Nature at Its Joints*. J. K. Campbell, M. O'Rourke, and M. H. Slater, eds. Cambridge, MA: MIT Press.

Willis, B. H., Beebee, H., and Lasserson, D. S. (2013), Philosophy of Science and the Diagnostic Process, *Family Practice*, 30:501–505.

World Health Organization (1992), *International Classification of Diseases*, 10th revision, Geneva: World Health Organization.

World Psychiatric Association (2003), Essentials of the World Psychiatric Association's International Guidelines for Diagnostic Assessment (IGDA), *British Journal of Psychiatry*, 182:S37–S66.

Wright, B., Dave, S., and Dogra, N. (2010), *100 Cases in Psychiatry*, Boca Raton, FL: CRC Press.

26

Mental Health and Illness: Past Debates and Future Directions

Reinier Schuur

Psychiatry is often taken to represent the application of medical science and practice to the human mind. Modern medicine has clearly done a lot to alleviate all manner of suffering from physical ailments that were once a normal part of life.[1] It has also helped reduce the stigma of certain physical conditions that were once thought to be signs of a weak moral character for which one is responsible. But what exactly the application of medicine to the realm of the human mind entails is not always so clear. To the extent that medicine concerns itself with the treatment of illness, it implies that the mind can be ill in some sense, that there can be such a thing as *mental* illnesses. Psychiatry is therefore immediately confronted by questions of a conceptual nature. What exactly would it even *mean* for the mind to be "ill"? How do mental illnesses relate to other problems that we face in life, such as struggling with loss, or perhaps our own character flaws? How does the concept of mental illness fit in with our conception of ourselves as free agents who make decisions for which we are responsible? These are questions of philosophy, which deals with clarifying how we *conceive* of ourselves and our place in the world, and developing our understanding of this fundamental relationship. Psychiatry in particular raises the following questions: Do we have any psychological capacity to *change* ourselves, what would that even mean, what would constitute a change for the better, and what should we not change but accept as healthy forms of suffering, like grief? If psychiatry is the application of medicine to the human mind, one concern is that we will address these questions of who we are and how we should live our lives with biomedical answers, that we will attempt to alleviate all forms of suffering, such as grief, as diseases to be treated, and deny any and all failures of character and responsibility. This would threaten our self-conception as free agents who have

some responsibility for who we are and can change ourselves for the better, and the idea that some struggles are a normal or even valuable part of life. This raises a question that has been the subject of much debate, namely, what do various medical concepts, such as health, disease, and illness, even mean in general, and are we justified or not in applying those concepts to the things that we call mental illnesses in mental health sciences and practices?

This chapter presents the past answers to this question and the future directions for this debate. In the section "Origins, assumptions, and main positions," I present the origins of this debate with Thomas Szasz's criticism of psychiatry. This includes explaining two widely held assumptions in the literature: the validity of the traditional philosophical method of conceptual analysis and the fact/value distinction. In the section "Responses to Szasz's conceptual analysis," I present the three reactions to Szasz's conceptual analysis: Naturalism, Normativism, and Hybrid accounts. In the section "Future directions," I show how the two mentioned assumptions of the existing literature can no longer be taken for granted given the current state of philosophy, and that taking these developments seriously shows promise in advancing a deeper understanding of mental health and illness.

Origins, assumptions, and main positions

The contemporary debate about the concept of mental illness and illness in general can be traced back to Thomas Szasz's criticism of psychiatry in the 1960s[2] (Szasz 1961). He argued that psychiatry not only is used as a tool for social-control,[3] but that it represents the misapplication of medical concepts to what he called normal "problems of living." Psychiatry certainly does continue to face a lot of criticism that it *medicalizes* "normal suffering."[4] But Szasz held that most, if not all, the conditions psychiatry treats, such as schizophrenia, are problems of living. The concept of "mental illness" does not refer to actual illnesses at all. Mental illness is a myth. Szasz did not deny that people had schizophrenia or could suffer from it. But what then did he mean? This requires us to understand what Szasz took the concept of illness to mean to determine what we can apply it to.

But if someone means one thing by illness, such disabilities with physical causes like cancer, and I mean any kind of disability, such as severe depression in response to life events, this just seems to be an issue of semantics. Different people can mean different things by the same concept. But what we really care

about is which meaning of the concept of illness we *should* use to settle which conditions to apply them to, such as depression. Philosophers disagree about how we should use concepts. This is because it depends on what we take concepts to be. This debate holds that (1) we do share concepts after all that (2) already have some basis in reality and (3) have necessary and sufficient conditions. In a certain sense, we do share medical concepts such as illness that already have a basis in reality, since we use them to refer to living things. What we mean by "sick" jokes is different from a sick person. But this doesn't tell us which specific definition of illness we should use to settle which conditions to apply it to. But if we assume that concepts have implicit necessary and sufficient conditions, then we can. A common analogy that is often used to explain this is that we can often tell whether or not a sentence is grammatical without being aware of the grammatical rule that explains this. The method by which we extract such a concept's implicit rules of application is called a conceptual analysis, where we draw on our intuitions to counter-examples of a concept and its relation to other concepts. Saying that illnesses are just conditions that warrant medical care would mean pregnancy is an illness. But since we do not apply the concept of illness to pregnancy, this shows us that medical care is not a necessary or sufficient condition for something being an illness. Clarifying what we mean by illness will therefore tell us what illnesses really are and whether we should apply them to what we call mental illnesses.

Szasz's analysis

The concept of physical illness for Szasz (1961) necessarily implies it having a physical cause. A physical cause in turn entails that an illness occurs "passively." Indeed, it seems that one thing that all physical illnesses have in common is that they are things that "happen to us," not things "we do" that we have control over. No one "chooses" to have cancer; it is something that just happens to us.[5] Szasz also held that the concept of physical illness implies that there is something wrong with the natural functioning of our bodies. One of the natural functions of the cell is to be integrated with the rest of our bodily functions. When cells become autonomous, we have cancer. Such biological norms are given by nature; they exist independently of how we choose to evaluate them. How we evaluate cancer is up to us, but it is counter-intuitive that we dictate the natural functions of cells.

Szasz's conceptual analysis therefore claims that illness implies passivity and natural norms. But for Szasz, these features do not apply to the things

we call "mental illnesses." Rather, mental illnesses are things that do not just happen to us, like cancer, they are things where *what we do* forms a central part of the condition's onset, development, and/or treatment. What exactly Szasz thought that we do or do not have choice over in mental illness is not always clear, and whether we have free will, as he thought, is heavily disputed and often dismissed as unscientific. But consider the fact that we would consider it ineffective to use psychotherapy to *treat* dementia. The reason is because we assume that dementia is not something that can be treated through acceptance, changing maladaptive ways of thinking, or reevaluating our circumstances, while we do think this is appropriate for conditions like depression following the loss of one's child. The implicit assumption here is that physical illnesses like dementia are not subject to change with the use of our minds over time, whereas mental illnesses like depression in a certain sense are; the concept mental illness for Szasz denies that we retain agency over how we use our own minds.[6]

Because mental illnesses involve our role as active agents and the choices we make, for Szasz mental illnesses are subject to ethical norms for how we ought to act. But unlike physical illnesses, ethical norms are not given by nature. Rather, ethical norms are provided by one's society. Such norms are much like the rules of a game, and playing games well for Szasz means playing the social roles that society provides (such as mother, ruler, teacher) well and living up to our moral duties. However, playing a socially valued role for Szasz is only a necessary but not a sufficient condition for what we call mental health. For mental health it is sufficient that the role one plays is self-selected. This also requires that one develop certain "meta-values," such as autonomy and self-knowledge, in order to pursue the particular roles that one individually values.

Because rules of ethics, according to Szasz, are provided by society, his distinction between physical and mental illnesses seems to be between biological versus social norms. But while Szasz did hold this view too, this is not the essential distinction. What is essential is that the social norms for how *we* as free agents ethically ought to act are "subjective," whereas natural norms for how the *body* ought to be are based on objective facts about our biology (or even natural facts about our own psychology). What exactly is meant by objective versus subjective norms is not always so clear in this debate. Objective norms are often taken to mean universal norms for how we ought to be or live, denying that there are different social or individual values for how we ought to live, which is what many mean by subjective norms. Indeed, what is a good

life for me might mean choosing a certain career, whereas a different career might make you happy. And there is no objective basis to say whether one career choice is better than the other if it makes us happy. But no one really denies that there are such individual differences. Indeed, it seems to be an empirical objective fact that a certain kind of career will make you happy whereas the same career would not make me happy. But there is an underlying assumption here. The assumption is that a life of happiness and free of needless suffering is indeed a good life and that this is what we ought to act for. But on what basis do we justify this assumption? Why ought we to act for our own happiness as opposed to living a life of suffering, perhaps because we think this will please God or serve society? If the claim that we ought to strive for some kind of life is some fundamental normative claim, what non-moral or non-circular claim can we appeal to justify living a certain way? Perhaps we cannot say why one ought to strive for well-being or suffering any more than chocolate- instead of vanilla-flavored ice cream.

The problem here is Hume's fact/value distinction, which holds that we cannot derive any normative claims for how we ought to live or act from objective facts without such normative claims already being assumed by our factual claims. For example, we may argue that, because animals suffer (factual claim) we ought not to eat them (normative claim). But the normative claim only follows because we are already assumed that suffering is bad in our factual claim. This is begging the question. A common conclusion that is drawn from this is that ethics is ultimately arbitrary, just a matter of preference. Indeed, if ethical norms are playing a game well by following the rules, then it follows that ethics is arbitrary since the rules of a game are simply agreed upon and therefore also arbitrary.

The view that ethics is arbitrary was widely held at this time and has also been attributed to Szasz. This is why Szasz rejects the concept of mental illness. It already assumes that since there are certain facts about our biology and psychology that explain impairments and suffering, these are objectively bad ways to live life. But since we cannot infer normative claims from such facts, mental health sciences and practices falsely assume that how we ought to live is based on objective norms. It is therefore irrelevant whether the suffering is normal, such as grieving, or severe, like schizophrenia. Since mental illnesses involve us as agents, mental health science and practice is fundamentally ethical and therefore subjective. The claim that mental illnesses are real illnesses denies this; it implies there are objective ethical norms, which for Szasz is a myth.

Responses to Szasz's conceptual analysis

Normativism

In response to Szasz, some have argued that physical illnesses are just as much based on subjective values as are mental illnesses, not objective norms of how the body ought to function. Cancer would then not be an illness. Nothing would be inherently bad about autonomous and destructive cells; cancer is just a certain state of the body that we don't like. This is simply pushing the fact/value distinction to its logical conclusion to include biological norms, not just ethical norms. Those who hold that illnesses in general are based on non-objective, and therefore, subjective norms are called Normativists (Engelhardt 1976; Fulford 1989; Goosen 1980; Sedgwick 1973).

But even if we allow for objective natural norms, like increased survival and reproduction, and make the factual claim that medical illnesses are reductions in biological fitness, we cannot infer from this that we ought to value biological fitness. This would violate the fact/value distinction, since the inference already assumes that we ought to value biological fitness and that reductions in biological fitness are objectively bad for us. But this is begging the question. This suggests that reductions in biological fitness are simply not what we mean by medical illness. Sickle cell disease promotes overall biological fitness by offering increased immunity to malaria, but we still see it as a medical condition. We seek medical care not to increase our biological fitness to promulgate the species. Rather, we seek medical care to extend our lives, have children, and relieve ourselves of needless suffering, because we tend to value these things in life. As Goosen points out, "Medicine serves not the species nor the individual's lineage, but the individual patient" (1980: 113).

But if cancer is an illness only because we happen to value living longer lives free of suffering, then illnesses are just anything about our bodies that we do not like. But this cannot be what we mean by illness either since there are other things we may not like about ourselves or our bodies that we do not see as illnesses, like being short or bald. Some Normativists, like Szasz, argue that illnesses are conditions that impair certain meta-values, such as autonomous, rational ordinary action, that allow us to pursue our subjective values. But now we face a problem. If we define autonomy, rationality, and ordinary actions too loosely, then we cannot rule out shortness because someone might see it as impeding ordinary action that they value. Indeed, shortness is seen as a major job liability in China, but not as an illness. However, if we argue that in this case it is irrational, that it does not impede autonomy, then we are being too specific,

and implying that there are things we should and should not objectively value, it is not subjective after all. So while meta-values seem to be important to what we think illnesses are, they do not seem to explain what we really mean by illnesses because it fails to explain such distinctions we make between different subjective values. Another route would be to draw on certain facts about us. Indeed, it can't be a coincidence that we tend to value things that also overlap with reductions in biological fitness. We therefore want to figure out what this relationship is. Finding such facts does not violate the fact/value distinction, it just tells us what distinguishes health values from other things we value.

Naturalism

So far our conceptual analysis has shown that illness involves subjective values but that they are not sufficient to explain what we mean by illness. Those who claim that it is certain facts about us that clarify the meaning of illness as distinct from other things we don't value are called Naturalists.

The most famous Naturalist account is Christopher Boorse's bio-statistical theory (1977). While Naturalists like Boorse do not deny that illnesses are the subject of our negative evaluations, they claim that the concept of illness has a basis in objective facts about us that are completely independent of our evaluations of them. Illness has this basis in the concept of disease, a theoretical concept separated from medical practice. His official definition of disease is quite technical,[7] but in essence for Boorse a disease is a reduction in present rates of biological fitness. However, as we saw, a reduction in biological fitness cannot be sufficient for something being an illness since this would violate the fact/value distinction. But for Boorse, what is also required for something to be an illness is that it is a disease that is not valued. This may appear to violate the fact/value distinction, but it is not intended to, since it is not claimed that we ought to value biological fitness as such. Rather, a disease being a necessary condition for illness only explains what kind of values illness refers to, namely, biological fitness rather than aesthetics values like eye color. We only ought to value biological fitness if the consequences of a reduction in biological fitness, such as with cancer, is something we do not like. Illnesses are those reductions in biological fitness that we do not value having, and this explains what we mean by illness.

The two main criticisms of Boorse's account are that there are counter-examples to his conceptual analysis and that his analysis of disease turns out to violate the fact/value distinction. While diseases as reductions in biological fitness rule out baldness and other things we don't like about ourselves as

distinct from illnesses, this definition is now too narrow and rules out too much, such as sickle cell disease which increases overall biological fitness. And even if all diseases involved a reduction in biological fitness, this already seems to assume that only the conditions that reduce our biological fitness ought to be valued. But this seems to assume that there are objectively better ways to live, namely, that we ought to not like those conditions that reduce biological fitness. But this cannot be assumed, since it begs the question whether that is what we should do. Medicinal science and practice does after all then seem to assume fundamentally ethical norms. And since we all accept in this debate that ethical norms are subjective, then medical science and practice is also inherently subjective. Disease as objective is a myth.

Hybrid accounts

The third conceptual analysis in this debate tries to ground facts about us and their consequences on our values by combining Naturalism and Normativism into one concept, so-called Hybrid accounts. The most well-known account is Wakefield's harmful dysfunction analysis (HDA) (1992).

Around the late 1970s, the concept of mental illness was being replaced with the concept of mental disorder. The concept of disorder is supposed to be equivalent to Boorse's concept of disease. But Wakefield does not distinguish between medical science as value-free and medical practice as value-laden. Medical science studies the conditions that medical practice treats, and so is equally determined by our social values. But Wakefield argues that the concept of disorder also has some basis in objective facts about us. He grounds the concept of disorder in dysfunctions, but he has a different analysis of function than Boorse does. Rather than being failures to contribute to present levels of biological fitness, dysfunctions are failures of a mechanism to perform the function they were selected for in our evolutionary past. However, this notion of dysfunction is only explanatory for why we think a mechanism is "broken." We still need a consequentialist concept in our analysis if we were to make dysfunctions relevant to what we value in medical practice. This is what Wakefield's concept of harm does. A disorder is a dysfunction that causes harm to someone, a harmful dysfunction. Indeed, we consider conditions to be medical illnesses because of their potential to cause ourselves harm. In turn, the concept of disorder constrains the relevant harms to those that are caused by dysfunctions, distinguishing disorders from other harms, like poverty or being not physically appealing. And since there are mental dysfunctions, those that

cause harm are mental disorders. So mental disorders can exist, contra Szasz, and are grounded in both facts and values, while the values are still not derived from such facts, only constrained by them.

The HDA does get closer to grounding the concept of disorder. It is harmful consequences of the breakdown of our bodies and minds, not their failure to contribute to fitness, that makes something a disorder. Indeed, a failure to function as designed by our evolutionary past does cause harm, such as the heart's failure to pump blood in cardiac arrest. However, the HDA faces several problems. First, it is empirically doubtful that all mental disorders are also evolutionary dysfunctions. But even if they happened to be such dysfunctions, this cannot be what makes them medical disorders. If conditions like phobias or depression happened to be selected for, we would still see them as disorders because of their harm. It remains unclear why only harms that are caused by evolutionary dysfunctions would make something a disorder but not other harms, like environmental mismatches. But the deeper reason why traditional conceptual analyses have not been successful in grounding the concept of disorder is because if we maintain the fact/value distinction, that part of mental health norms is purely contingent on the subject, that is, subjective, then no necessary conditions can ground the concept of disorder in objective facts. Values' contingent relationships to facts cannot be made necessary by stipulation. We cannot try to bridge the fact/value distinction while simultaneously trying to keep facts and values apart.

Future directions

The debate over the concept of mental illness has been going on since the 1960s and has assumed the validity of traditional conceptual analysis and the fact/value distinction. But these assumptions have been challenged by contemporary philosophy. Despite this, the concept of mental illness debate has remained relatively isolated from these developments. In this second half of the chapter I will briefly show how bringing contemporary philosophy to bear on this debate has promise to move it forward and make it more relevant to mental health science and practice.

Traditional conceptual analysis

A generic conceptual analysis is indispensable to fundamental questions about the nature of health and illness, physical or mental, such as showing us that what

is at issue are which norms and what role passivity and agency play in having an illness. However, the use of traditional conceptual analysis in this debate assumes a lot more, such as the classical theory of concepts as having necessary and sufficient conditions that exist independently of us that we all share in some sense, and that we can access through our reflections and intuitions. For example, Szasz took his conceptual analysis to pick out a relatively clear distinction between physical and mental disorders that our concept of mental illness corresponds to. However, such a strong distinction has been rejected by contemporary empirical sciences of the mind. Our concept of mental versus physical is inherently confused, and does not look to be a natural kind (Murphy 2006). Moreover, traditional conceptual analysis and its assumptions have been called into question, especially in the last five years or so, starting a larger debate about philosophical methodology (Strevens, forthcoming; Machery 2017). For example, empirical psychology has challenged the classical theory of concepts by showing that their application does not have necessary and sufficient conditions, and alternative theories have been proposed to explain the application of concepts, such as prototype theories. Moreover, the universal nature of concepts as shared by everyone has been criticized by showing how different social groups have different intuitions about the application of certain concepts (Machery 2017). All of this challenges the idea that our concepts are already "out there" reflecting reality, waiting for us to clarify them by analyzing them. We can see this in this debate in particular in the assumption that there is a clear distinction between objective norms found in the natural sciences from social norms of mental health. This brings us to the fact/value distinction.

Facts and values

What exactly the fact/value distinction means is not always so clear. In this debate, it seems to mean that when it comes to ethical and health norms, we cannot appeal to objective facts alone to say which are better or worse. Such norms are matter of social custom or personal preference, which is what we mean by "subjective" rather than "objective." But while these terms are equally unclear, their purported acceptance has legitimate motivations that need to be acknowledged.

First, there *are* differences in values between both social groups and individuals in how to live a good life where we may indeed not be able to say which are better or worse, such as individual values concerning occupational choices and different artistic expressions in cultures. But the idea that there are

objective norms is historically often taken to imply that there is one and only one way to live a good life, such as the life of the philosopher according to Aristotle. To the extent that we have been wrong in the past about what constitutes a good life, like being Christian and straight, and that there are a multitude of ways to living a good life, we may therefore want to reject objective ethical and mental health norms. Second, claims about objective norms have historically been used as the basis to abuse and oppress individuals or groups of people that deviate from those norms. We therefore have legitimate political concerns about claims of objective norms.

However, the political concern is distinct from the existence of objective and subjective norms, and has been dealt with by more constructive means than accepting subjectivism. As for the fact/value distinction itself, when we ask for concrete examples to clarify it and attempt to apply it to its logical conclusion, problems begin to emerge in our conceptual analysis of mental illness. This is because it is simply not true that all that we mean by mental illnesses are just deviations from our social agreements and personal preferences, that disease is equivalent to other social violations, like failures of etiquette. That's just not what we mean by the concept. Even if it is subjective norms that ground the concept of mental illness, it fails as a conceptual analysis to make distinguish between mental health norms and other subjective norms, such as other negative evaluations of mental traits and dispositions that we do not like, such as being really annoying.

The solution sought for this "run-away" subjectivism by all positions, including, Szasz's account, Normativism, Naturalism, and Hybrid add constraints in their conceptual analyses as a necessary basis for mental health and illness, such as "meta-values" that makes the pursuit of specific values possible, or objective norms such as evolutionary norms. But to the extent that such constraints are supposed to limit such cases, they seem incompatible with the fact/value distinction. This is because they are based on certain facts about us as rational and social animals with various requirements in order to function well based on how we use such capacities, such as acquiring knowledge of reality and ourselves and cultivating autonomy. That there are both facts about us that we tend to share with other animals and most human beings, as well as unique individual differences that determine mental health, is clear. The question now is what those shared facts and individual differences are. But at this point what exactly it means for norms to be objective or subjective is even less clear than before. We need to clarify what we mean by *these* concepts if we are to move this debate forward rather than assuming that we know what we are talking about.

There is currently a debate in ethics that deals specifically with just this question of what objective versus subjective factors constitute our psychological well-being (Tiberius and Plakias 2010). This debate is part of a larger development in ethics since the 1970s, a revival of virtue ethics in the ancient Greek philosophical tradition (Anscombe 1958; Foot 2001). These developments also call into question the fact/value distinction by arguing that what we ought to do to live good lives can be determined by what human beings require to flourish based on objective facts about human nature, such as our biology and psychology. Moreover, unlike previous objective accounts of the good life, these accounts of well-being attempt to accommodate a higher diversity of ways of living good lives (Nussbaum 2007), while still adding certain objective standards, such as dignity, much like the Normativists' meta-values in the mental illness debate.

This debate has also led to research on the relationships between virtues, such as self-control, emotional regulation, resilience, and their relationship to mental health. The fruits of this work are new frameworks for therapeutic interventions that can help empower and bolster the agency of patients, making them more active in their own care (Pearce and Pickard 2010). This requires us to fundamentally rethink underlying distinctions that psychiatry and those fighting mental illness stigma make, such as that mental illness is "chemical, not character," and cannot be both. So Szasz may have been right that some mental illnesses have a moral component, but wrong that this makes them mutually exclusive to also being medical disorders with biological components.

An important feature of these developments is its interaction with the empirical sciences, such as psychological work on theories of well-being, character-development, and self-control (Pickard 2013). Such work however, still requires some form of conceptual analysis. Analysis can help us think more clearly about our underlying conceptual assumptions that determine how we interpret empirical data and relate it to our larger conceptual scheme, as opposed to accepting or leaving unclear how we think of objective and subjective norms. But as we have seen, traditional conceptual analysis tends to assume that most of our concepts already tend to map onto clear distinctions in reality. To the extent that this method has been called into question by philosophy, it is no longer adequate if we want an empirically informed debate about mental health and illness.

In response to the criticisms of traditional conceptual analysis, a debate is going on about alternative philosophical methods that are more empirically informed and practically oriented, such as philosophical explication is the tradition of Rudolf Carnap, which focuses not primarily on what we mean by concepts, but also how

to improve the role that they play in science. Another example is conceptual engineering, which seeks to address the broader question of what constitutes having a good concept and their uses in various fields, such as philosophy, policy, and science (Eklund 2015). Drawing on such research will help us develop better standards for clear conceptual thinking that mental health sciences and practices need, such as developing our understanding of the relationship between physical and mental illnesses, and morality and mental health.

Conclusion

The concept of mental illness debate was a response to legitimate concerns regarding abuses of psychiatry and the medicalization of normal suffering and differences. Traditional conceptual analysis assumes that our underlying conceptual scheme that our common sense intuitions and existing medical expertise are based on can tell us what mental illness is. However, whether they can is precisely what is at issue, it cannot be assumed. In particular, our assumption of the fact/value distinction can no longer be taken for granted given contemporary debates in ethics. And if the issue is about general norms versus individual differences in mental health, this is something that science can objectively study. But to the extent that this requires us to rethink our assumptions, traditional conceptual analysis is problematic since it assumes that our underlying conceptual scheme only needs to be clarified as opposed to being challenged and developed. Drawing on current philosophical methods that are more empirical and practically oriented will therefore help make the concept of mental health debate relevant to mental health science and practice.

Notes

1 Tooth decay used to be rife, now it is a disease. Improving standards changed what it means to be healthy.

2 Szasz was part of the anti-psychiatry movement, a label he rejected, which was a reaction to the history of involuntary treatment and other abuses by psychiatry. Other figures include Michel Foucault and R.D. Laing.

3 For example, in nineteenth-century America, drapetomania was the "condition" that caused black slaves to run away from their masters, and political dissent was labeled a "mental illness" in the USSR (Szasz 1961).

4 Mourning, it has been argued, can now be diagnosed as major depressive disorder in the *DSM*-5, since it no longer rules out depression in response to the loss of a loved one (Horwitz and Wakefield 2007).
5 For Szasz, the passivity of illness also implies that medical interventions, such as surgery and medication, are appropriate forms of treatment, since illnesses are not subject to one's direct will to change long term.
6 For this reason then, medication and surgery are inappropriate treatments for mental illnesses for Szasz.
7 Boorse defines disease as a failure to contribute to statically typical levels of biological fitness according to one's reference class.

References

Anscombe, G. E. M. (1958), Modern Moral Philosophy, *Philosophy*, 33:1–19.
Boorse, C. (1977), Health as a Theoretical Concept, *Philosophy of Science*, 44:542–573.
Eklund, M. (2015), Intuitions, Conceptual Engineering, and Conceptual Fixed Points. In *The Palgrave Handbook of Philosophical Methods*. C. Daly, ed. London: Palgrave Macmillan.
Engelhardt, H. T. (1976), Ideology and Etiology, *The Journal of Medicine and Philosophy*, 1:256–268.
Foot, P. (2001), *Natural Goodness*, Oxford: Oxford University Press.
Fulford, K. W. M. (1989), *Moral Theory and Medical Practice*, Cambridge: Cambridge University Press.
Goosens, W. K. (1980), Values, Health, and Medicine, *Philosophy of Science*, 47:100–115.
Horwitz, A. V. and Wakefield, J. C. (2007), *The Loss of Sadness: How Psychiatry Transformed Normal Sorrow into Depressive Disorder*, Oxford: Oxford University Press.
Machery, E. (2017), *Philosophy within Its Proper Bounds*, Oxford: Oxford University Press.
Murphy, D. (2006), *Psychiatry in the Scientific Image*, Cambridge, MA: MIT Press.
Nussbaum, M. C. (2007), *Frontiers of Justice: Disability, Nationality, Species Membership: The Tanner Lectures on Human Values*, Cambridge, MA: Belknap Press, Harvard University Press.
Pearce, S. and Pickard, H. (2010), Finding the Will to Recover: Philosophical Perspectives on Agency and the Sick Role, *Journal of Medical Ethics*, 36:831–833.
Pickard, H. (2013), Psychopathology and the Ability to Do Otherwise, *Philosophy and Phenomenological Research*, 90:135–163.
Sedgwick, P. (1973), Illness: Mental and Otherwise, *The Hastings Center Studies*, 1:19–40.

Szasz, T. S. (1961), *The Myth of Mental Illness: Foundations of a Theory of Personal Conduct*, New York: Harper Collins.

Tiberius, V. and Plakias A. (2010), Well-Being, pp. 402–432. In *The Moral Psychology Handbook*. J. M. Doris and The Moral Psychology Research Group, eds. Oxford: Oxford University Press.

Wakefield, J. C. (1992), The Concept of Mental Disorder: On the Boundaries between Biological Facts and Social Values, *American Psychologist*, 47:373–388.

Study Questions

1. *What is mental disorder?*
2. *What is the difference between diagnosis in psychiatry and medicine?*
3. *Should psychiatry adopt reductionism or holism as an ideal?*
4. *What is the role of gender and race in psychiatry?*
5. *How do social institutions affect the quality of mental health care?*

Annotated Bibliography/Further Reading

Ankeny, R. (2017), The Role of Patient Perspectives in Clinical Case Reporting, pp. 97–112. In *Knowing and Acting in Medicine*. R. Bluhm, ed. New York: Rowman and Littlefield.

Provides an analysis of the role patient perspectives in clinical case reporting in medicine.

Appelbaum, P. S. (2004), Psychiatric Advance Directives and the Treatment of Committed Patients, *Psychiatric Services*, 55:751–763.

Presents information on *advance directives*, one of the promising innovations to give *patients* a greater voice in their *psychiatric treatment*.

American Psychiatric Association (2013), *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Arlington, VA: American Psychiatric Publishing.

Diagnostic manual that offers a common language and standard criteria for the classification of mental disorders. It is used, or relied upon, by clinicians, researchers, drug regulation agencies, health insurance companies, pharmaceutical companies, the legal system, and policymakers.

Appelbaum, P. S. and Gutheil, T. G. (2007), *Clinical Handbook of Psychiatry and the Law*, 4th edition. Philadelphia, PA: Wolters Kluwer.

Exposes the critical issues that take place at the intersection of psychiatry and the law.

Berrios, G. E. (1996), *The History of Mental Symptoms: Descriptive Psychopathology since the Nineteenth Century*, Cambridge: Cambridge University Press.

Historical and philosophical evaluation of descriptive models of psychopathology.

Bluhm, R. (2009), Gender Differences in Depression: Explanations from Feminist Ethics, *International Journal of Feminist Approaches to Bioethics*, 4(1):69–88.

Uses feminist analyses of personhood and of autonomy to explain the greater prevalence of depression among women.

Boorse, C. (1975), On the Distinction between Disease and Illness, *Philosophy and Public Affairs*, 5:49–68.

Develops a philosophical approach to distinguish disease from illness.

Boorse, C. (1977), Health as a Theoretical Concept, *Philosophy of Science*, 44:542–573.

Develops a conceptual approach to defining health.

Borsboom, D. (2017), A Network Theory of Mental Disorders, *World Psychiatry*, 16:5–13.

An overview of the network approach to understanding mental disorder.

Borsboom, D. and Cramer, A. O. J. (2013), Network Analysis: An Integrative Approach to the Structure of Psychopathology, *Annual Review of Clinical Psychology*, 9:91–121.
Suggests that psychiatric disorders may be best understood as self-sustaining causal networks of symptoms and presents methods for identifying these networks.

Broome, M. and Bortolotti L., eds. (2009), *Psychiatry as Cognitive Neuroscience*, Oxford: Oxford University Press.
Analyzes the scope and limitations of neuroscience/cognitive neuroscience in understanding mental illness with a focus on a broad range of disorders, such as depression, schizophrenia, personality disorders, and a wide range of neuroscientific methods, such as functional neuroimaging and neuropsychological models.

Bolton, D. (2008), *What Is Mental Disorder? An Essay in Philosophy, Science, and Values*, Oxford: Oxford University Press.
Develops a pragmatic definition of mental disorder through an evaluation of how we diagnose, define, and treat it.

Boyd R. (1991), Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds, *Philosophical Studies*, 61:117–148.
Develops a historical and philosophical evaluation of the natural kind debate in philosophy of science.

Bracken, P. Thomas P. (2010), From Szasz to Foucault: On the Role of Critical Psychiatry, *Philosophy, Psychiatry, and Psychology*, 17(3):219–228.
Defence of the necessity of the critical approaches to psychiatry through a historical analysis.

Capponi, P. (1992), *Upstairs in the Crazy House*, Toronto: Viking.
An account of how Canada's reduced social programs affect mental illness treatment with a focus on the unique communities the poor create, how poverty differs from city to city, and the inventive ways some groups have found to improve their conditions and lives.

Carel, H. and Kidd, I. J. (2014), Epistemic Injustice in Healthcare: A Philosophical Analysis, *Medicine, Health Care, and Philosophy*, 17:529–540.
A critical evaluation of epistemic injustice in healthcare.

Cartwright, S. A. (1851/2004), Diseases and Physical Peculiarities of the Negro Race, pp. 28–39. In *Health, Disease, and Illness*. A. L. Caplan, J. J. McCartney, and D. A. Sisti, eds. Washington, DC: Georgetown University Press.
An article that construes the attempt to run away from slavery as a form of mental illness.

Charland, L. C. (2004), A Madness for Identity: Psychiatric Labels, Consumer Autonomy, and the Perils of the Internet, *Philosophy, Psychiatry, and Psychology*, 11(4):335–349.
Discusses how a diagnosis of a mental disorder can influence peoples' self-concepts, particularly in the era of the internet.

Charland, L. C., Hope, T., and Tan J. (2013), Anorexia Nervosa as a Passion, *Philosophy, Psychiatry, and Psychology*, 20(4):v353–v365.
Draws on both philosophical analysis and patient interviews to reconceive anorexia nervosa.

Cheney, T. (2009), *Manic: A Memoir*, New York: HarperCollins Publishers.
A Beverly Hills Lawyer's recounting of her encounter with bipolar disorder.

Chung M. C., Fulford, K. W. M., and Graham G., eds. (2006), *Reconceiving Schizophrenia*, Oxford: Oxford University Press.
A collection of multiple perspectives on understanding schizophrenia.

Cooper, R. (2005), *Classifying Madness: A Philosophical Examination of the Diagnostic and Statistical Manual of Mental Disorders*, Dordrecht: Springer.
Examines the fundamental philosophical issues in the *Diagnostic and Statistical Manual of Mental Disorders*.

Cooper, R. (2007), *Psychiatry and Philosophy of Science*, Montreal and Kingston: McGill-Queen's University Press.
Introduces philosophy of psychiatry from the perspective of philosophy of science.

Corrigan, P. (2004), How Stigma Interferes with Mental Health, *American Psychologist*, 59:614–625.
Evaluates the negative impact of stigma on mental health.

Cosgrove, L., Krimsky, S., Vijayaraghavan, M., and Schneider, L. (2006), Financial Ties between DSM-IV Panel Members and the Pharmaceutical Industry, *Psychotherapy and Psychosomatics*, 75:154–160.
Exposes the conflicts of interests of the *DSM*-IV panel members.

Cosgrove, L. and Krimsky, S. (2012), A Comparison of DSM-IV and DSM-5 Panel Members' Financial Associations with Industry: A Pernicious Problem Persists, *PLoS Medicine*, 9:3.
Draws attention to the similarities between the conflicts of interests of the *DSM*-IV panel members with that of the *DSM*-5.

Cuthbert, B. N. (2014), The RDoC Framework: Facilitating Transition from ICD/DSM to Dimensional Approaches That Integrate Neuroscience and Psychopathology, *World Psychiatry*, 13(1):28–35.
Lays out the fundamental tenets of the RDoC framework through a comparison between the *ICD* and the *DSM*.

Cuthbert, B. N. and Insel, T. R. (2013), Toward the Future of Psychiatric Diagnosis: The Seven Pillars of RDoC, *BMC Medicine*, 11(126). URL= http://www.biomedcentral.com/1741-7015/11/126
Lays out the fundamental properties of the National Institute of Mental Health's Research Domain Criteria, which was developed as an alternative to the *DSM*-5's framework for mental disorders.

Demazeux, S. and Singy, P., eds. (2015), *The DSM-5 in Perspective: Philosophical Reflections on the Psychiatric Babel*, Dordrecht: Springer.
The book is conceived as a direct response to the publication of *DSM*-5 and explores a wide range of issues and controversies related to both the process of developing the *DSM*-5, the changes that were affected, and the broad significance of the revised classification scheme.

Dumit, J. (2004), *Picturing Personhood: Brain Scans and Biomedical Identity*, Princeton, NJ: Princeton University Press.
Discusses the implications of the advances in neuroimaging to the conceptions of personhood.

Engel, G. L. (1977), The Need for a New Medical Model: A Challenge for Biomedicine, *Science*, 196:129–136.
Develops the influential biopsychosocial model of medicine.

Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G., and Munoz, R. (1972), Diagnostic Criteria for Use in Psychiatric Research, *Archives of General Psychiatry*, 26(1):57–63.
Develops a scientific criteria for diagnosing mental disorders through a measurement of symptoms and signs.

Foucault, M. (1965), *Madness and Civilization: A History of Insanity in the Age of Reason*, R. Howard, trans. New York: Pantheon.
Examines the archaeology of madness in the West from 1500 to 1800.

Freud, S. (1956–1974), *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. James Strachey, Trans. London: Hogarth Press, 24 volumes.
Freud's collected works.

Flanagan, O. (2013), Identity and Addiction: What Alcoholic Memoirs Teach, pp. 865–888. In *The Oxford Handbook of Philosophy and Psychiatry*. K. W. M. Fulford, Martin Davies, Richard G. T. Gipps, George Graham, John. Z. Sadler, Giovanni Stanghellini, and Tim Thornton, eds. Oxford: Oxford University Press.
A philosophical exploration of how memoirs of addiction may provide resources those battling with addiction.

Fulford, K. W. M., Davies, M., Gipps, R., Graham, G., Sadler, J., Stanghellini, G., and Thornton, T. (2013), *The Oxford Handbook of Philosophy and Psychiatry*, Oxford: Oxford University Press.
Offers a comprehensive review of the major progress in the field of philosophy of psychiatry, in primarily the areas of philosophy of mind, ethics, and philosophy of science.

Gardner, S. (2009), *Irrationality and the Philosophy of Psychoanalysis*, Cambridge: Cambridge University Press.
Argues that psychoanalytic theories provide the best way to understand the phenomenon of irrationality.

Gergel, T. and Kabir, T. (2017), Reframing a Model: The Benefits and Challenges of Service User Involvement in Mental Health Research, pp. 77–96. In *Knowing and Acting in Medicine*. Bluhm, R., ed. New York: Rowman and Littlefield.

Grünbaum, A. (1984), *The Foundations of Psychoanalysis: A Philosophical Critique*, Berkeley: University of California Press.
Provides an extensive criticism of the argument that psychoanalysis is a science.

Ghaemi, N. (2009), *The Rise and Fall of the Biopsychosocial Model: Reconciling Art and Science in Psychiatry*, Baltimore, MD: Johns Hopkins University Press.
Historical and philosophical study of the biopsychosocial model.

Ghaemi, N. (2012), *On Depression: Drugs, Diagnosis, and Despair in the Modern World*, Baltimore, MD: Johns Hopkins University Press.
A philosophical approach to depression in contemporary society.

Graham, G. and Stephens G. L. (1994), *When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts*, Cambridge MA: MIT Press.
Examines verbal hallucinations and thought insertion as examples of what the authors call alienated self-consciousness.

Grisso, T. and Appelbaum, P. S. (2001), *MacArthur Competence Assessment Tool for Treatment (MacCAT-T)*, Sarasota, FL: Professional Resource Press.
An assessment tool for evaluating competency of individuals with mental disorders.

Gupta, M. (2014), *Is Evidence-Based Psychiatry Ethical?* Oxford: Oxford University Press.
Evaluation of the ethical implications of evidence-based psychiatry.

Guze, S. B. (1992), *Why Psychiatry Is a Branch of Medicine*, Oxford: Oxford University Press.
Argues for a biomedical model of psychiatry.

Haslam, N. (2001), Kinds of Kinds: A Conceptual Taxonomy of Psychiatry Categories, *Philosophy, Psychiatry and Psychology*, 9(13):203–217.
Develops a taxonomy of categories in psychiatry.

Haslam, N. (2014), Natural Kinds in Psychiatry: Conceptually Implausible, Empirically Questionable, and Stigmatizing. In *Classifying Psychopathology*. H. Kincaid and J. A. Sullivan, eds. Massachusetts: MIT Press.
A critical approach to construal of mental disorders as natural kinds.

Hawthorne, S. C. C. (2013), *Accidental Intolerance: How We Stigmatize ADHD and How We Can Stop*, Oxford: Oxford University Press.
Shows how social values have influenced scientific and medical practices related to attention deficit hyperactivity disorder and outlines the consequences of this influence.

Hacking, I. (1995), The Looping Effects of Human Kinds, pp. 351–383. In *Causal Cognition: A Multidisciplinary Debate*. D. Sperber, D. Premack and A. J. Premack, eds. Oxford: Clarendon Press.
Contrasts human kinds (kinds of people, their behavior, their emotions, etc.), studied by social sciences, to natural kinds, studied by natural sciences, with the argument that the former is subject to looping effects.

Hacking, I. (1999), *The Social Construction of What?* Cambridge, MA: Harvard University Press.
Examines the social construction debate in philosophy by consideration of various examples that embody the inherent conflicts of the debate, such as child abuse and schizophrenia.

Hacking, I. (2007), Kinds of People: Moving Targets, *Proceedings of the British Academy*, 151:285–318.
Develops a detailed account of looping effects by considering how classifications made by human sciences interact with the very people who were classified, changing the properties of these people in return.

Horwitz, A. V. (2002), *Creating Mental Illness*, Chicago, IL: University of Chicago Press.
Argues that current conceptions of mental illness as a disease fit only a small number of serious psychological conditions and that most conditions currently regarded as mental illness are cultural constructions.

Herman, J. (1992), *Trauma and Recovery: The Aftermath of Abuse—From Domestic Abuse to Political Terror*, New York: Basic Books.
Groundbreaking work that provides a socially informed account of trauma.

Hoffman, G. (2016), Out of our Skulls: How the Extended Mind Thesis Can Extend Psychiatry, *Philosophical Psychiatry*, 29(8):1160–1174.
Application of the extended mind hypothesis to the understanding and treatment of mental disorders.

Hoffman, G. and Hansen J. L. (2011), Is Prozac a Feminist Drug? *International Journal of Feminist Approaches to Bioethics*, 4(1):89–120.
Develops a feminist evaluation of the use of antidepressants, such as Prozac.

Horwitz, A. V. and Wakefield, J. C. (2007), *The Loss of Sadness: How Psychiatry Transformed Normal Sorrow into Depressive Disorder*, New York: Oxford University Press.
A critical evaluation of the *DSM*'s construal of sadness and melancholy as depression.

James, W. (1894), What Is an Emotion? *Mind*, ox-IX(34):188–205.
A groundbreaking approach to understanding emotion in psychology.

Jamison, K. (1995), *An Unquiet Mind: Memoirs of Moods and Madness*, New York: Knopf.
A psychiatrist's first person account of her experience with bipolar disorder.

Jaspers, K. (1913/1997), *General Psychopathology*. J. Hoenig and M. W. Hamilton, trans. Baltimore, MD: Johns Hopkins University Press.
Key work in phenomenological psychopathology.

Kendell, R. E. and Jablensky, A. (2003), Distinguishing between the Validity and Utility of Psychiatric Diagnoses, *American Journal of Psychiatry*, 160:4–12.
Provides an overview of the use of "validity" in psychiatry and considers the question of whether psychiatric diagnoses are distinct entities.

Kendler, K. S. and Parnas, J., eds. (2012), *Philosophical Issues in Psychiatry II: Nosology*, Oxford: Oxford University Press.
This book was published during the revisions of the major psychiatric diagnostic manuals, the *DSM* and *ICD*. It evaluates major issues in psychiatric nosology in so far as it shapes psychiatric research, treatment, and insurance-related and administrative contexts.

Kendler, K. S. (1990), Toward a Scientific Psychiatric Nosology: Strengths and Limitations, *Archives of General Psychiatry*, 47:969–973.
A critical evaluation of the strengths and limits of contemporary psychiatric taxonomy.

Kendler, K. S. (2005), Towards a Philosophical Structure for Psychiatry, *American Journal of Psychiatry*, 162:434.
Defence of a philosophical approach to understanding and explaining mental disorders in psychiatry.

Kendler, K. S. (2008), Explanatory Models for Psychiatric Illness, *American Journal of Psychiatry*, 165:695–702.
Answers how we can best develop explanatory models for psychiatric disorders given that causal factors have an impact on psychiatric illness both within and outside of the individual, involving biological, psychological, and sociocultural processes.

Kendler, K. S., Zachar, P., and Craver, C. (2011), What Kinds of Things Are Psychiatric Disorders? *Psychological Medicine*, 41:1143–1150.
Evaluates whether mental disorders are natural kinds by adopting the homeostatic property cluster view of natural kinds.

Kincaid, H. and Sullivan J. A., eds. (2014), *Classifying Psychopathology: Mental Kinds and Natural Kinds*, Cambridge, MA: MIT Press.
Focuses on the question of whether current psychiatric classifications are close to meeting the scientific ideal of natural kinds, and the extent to which it is necessary for mental disorders to be labeled as natural kinds for them to be subject to rigorous scientific investigation.

Kleinman, A. (2004), Culture and Depression, *New England Journal of Medicine*, 351:951–953.
An anthropological approach to understanding the cultural influences on depression.

Kupfer, D. J., First, M. B., and Regier, D. A. (2002), *A Research Agenda for DSM-V*, Washington, DC: American Psychiatric Association, xvii.
Discusses the research priorities of the *DSM*-5 in its development process.

Kuhn, T. (1962), *The Structure of Scientific Revolutions*, Chicago, IL: University of Chicago Press.
Explains what science is and how scientific change occurs through a historical and philosophical approach.

Lane, C. (2007), *Shyness: How Normal Behavior Became a Sickness*, New Haven, CT: Yale University Press.
Historical account of the role of the *DSM* in the pathologization of normal behavior.

Lilienfeld, S. O. and Marino, L. (1995), Mental Disorder as a Roschian Concept: A Critique of Wakefield's "Harmful Dysfunction" Analysis, *Journal of Abnormal Psychology*, 104:411–420.
A philosophical criticism of the "harmful dysfunction" approach to defining mental disorder.

Luhrmann, T. M., Padmavati, R., Tharoor, H., and Osei, A. (2014), Differences in Voice-Hearing Experiences of People with Psychosis in the USA, India and Ghana: Interview-Based Study, *The British Journal of Psychiatry*, 206:41–44.
An anthropological comparison of voice-hearing experiences of individuals in the United States, India, and Ghana.

Maibom, H. L. (2008), The Mad, the Bad, and the Psychopath, *Neuroethics*, 1(3):167–184.
A philosophical approach to understanding psychopathy.

Maiese, M. (2016), *Embodied Selves and Divided Minds*, Oxford: Oxford University Press.
Focuses on the philosophical approaches to the self and personal identity and their connection to psychiatry. Arguing against the internalistic approaches to mind, it endorses the essentially embodied model of consciousness and cognition inspired by the works of existential phenomenologists, who emphasize the unique role of the human body in determining our mental states.

Merleau-Ponty, M. (1945/2012), *Phenomenology of Perception*. Donald Landes, trans. New York: Routledge.
The foundational work in phenomenology that describes the life-world of humans as perceiving subjects.

Moncrieff, J. (2010), Psychiatric Diagnosis as a Political Device, *Social Theory and Health*, 8:370–382.
Construal of psychiatric diagnosis as a tool for oppression.

Moseley, D. and Gala, G. (2015), *Philosophy and Psychiatry: Problems, Intersections and New Perspectives*, New York: Routledge.
The book fosters a dialogue between philosophers and psychiatrists, and as such, focuses primarily on issues related to diagnosis and treatment.

Murphy, D. (2006), *Psychiatry in the Scientific Image*, Cambridge, MA: MIT Press.
Evaluates psychiatry from the viewpoint of analytic philosophy of science, focusing on how mental illness should be conceived of, classified, and explained.

Nordenfelt, L. (1987), *On the Nature of Health: An Action-Theoretic Approach*, Dordrecht: Reidel.
Influential explication of the normativist approach to understanding health and disease.

Paris, J. and Phillips J., eds. (2013), *Making the DSM-5: Concepts and Controversies*, New York: Springer.
An interdisciplinary edited volume on the development of the 5th edition of the *DSM*.

Poland, J. and Graham, G., eds. (2011), *Addiction and Responsibility*, Cambridge MA: MIT Press.
Collection of critical essays on addiction and its implications for personhood and responsibility.

Potter, N. N. (2016), *The Virtue of Defiance and Psychiatric Engagement*, Oxford: Oxford University Press.
Discusses defiance in various social contexts.

Potter, N. N. (2009), *Mapping the Edges and the In-Between: A Critical Analysis of Borderline Personality Disorder*, Oxford: Oxford University Press.
Focuses on borderline personality disorder from the perspective of cultural and gender norms.

Radden, J. and Varga, S. (2013), The Epistemological Value of Depression Memoirs: A Meta-Analysis, pp. 99–116. In *The Oxford Handbook of Philosophy and Psychiatry*. K. W. M. Fulford, Martin Davies, Richard G. T. Gipps, George Graham, John Z. Sadler, Giovanni Stanghellini, and Tim Thornton eds. Oxford: Oxford University Press.
A philosophical inquiry into the epistemic value of patient memoirs.

Radden, J. (1996), *Divided Minds and Successive Selves*, Oxford: Oxford University Press.
Applies theories of personal identity to the complex cases of personality and identity disturbances in psychopathology and discusses the implications of such scenarios to ethical and legal issues.

Radden, J., ed. (2002), *The Nature of Melancholy: From Aristotle to Kristeva*, Oxford: Oxford University Press.
A historical and philosophical account of the evolution of melancholy over time.

Radden, J. (2004), *The Philosophy of Psychiatry: A Companion*, Oxford: Oxford University Press.
Collection of essays on various topics in philosophy of psychiatry.

Ratcliffe, M. (2008), *Feelings of Being: Phenomenology, Psychiatry, and the Sense of Reality*, Oxford: Oxford University Press.
Addresses the role of existential feelings in psychiatric conditions.

Sadler, J. Z. (2002), *Descriptions and Prescriptions: Values, Mental Disorders, and the DSMs*, Baltimore, MD: Johns Hopkins University Press.
Focuses on crucial philosophical, scientific, and ethical questions about psychiatric classifications.

Sadler, J. Z. (2005), *Values and Psychiatric Diagnosis*, Oxford: Oxford University Press.
A classic book that provides an in-depth examination of the methods of psychiatric diagnosis, both of the *DSM* and the *ICD*, by exploring specifically how psychiatric classifications are influenced by value judgments.

Sadock, B. J., and Sadock, V. A. (2011), *Kaplan and Sadock's Synopsis of Psychiatry*, 10th edition. USA: Lippincott Williams & Wilkins.
Central textbook in psychiatry.

Saks, E. (2007), *The Centre Cannot Hold: My Journey through Madness*, New York: Hyperion.
A successful law professor's recounting of her encounter with schizophrenia.

Szasz, T. (1961), *The Myth of Mental Illness: Foundations of a Theory of Personal Conduct*, New York: Harper & Row.
Defends the claim that mental illness is a myth.

Sullivan, J. A. (2016), Stabilizing Constructs through Collaboration across Different Research Fields as a Way to Foster the Integrative Approach of the Research Domain Criteria (RDoC) Project, *Frontiers in Human Neuroscience*, 10:309.
Highlights the necessity and the importance of construct stability in the neuroscientific approaches to mental disorders.

Tekin, Ş. (2014), Psychiatric Taxonomy: At the Crossroads of Science and Ethics, *Journal of Medical Ethics*, 40:513–514.
Evaluates the scientific classification of mental disorders from an ethical perspective.

Tekin, Ş. (2011), Self-Concept through the Diagnostic Looking Glass: Narratives and Mental Disorder, *Philosophical Psychology*, 24:357–380.
Discusses the impact of psychiatric diagnosis on personal identity and flourishing.

Tekin, Ş. (2014), Self-Insight in the Time of Mood Disorders: After the Diagnosis, Beyond the Treatment, *Philosophy, Psychiatry, and Psychology*, 21:139–155.
Focuses on the reflective impact of psychiatric diagnoses on self-insight.

Tekin, Ş. (2016), Are Mental Disorders Natural Kinds? A Plea for a New Approach to Intervention in Psychiatry, *Philosophy, Psychiatry, & Psychology*, 23(2):147–163.
Invites philosophers of psychiatry to move away from metaphysical discussions of whether mental disorders are natural kinds to epistemological discussions of scientific research on mental disorders.

Tekin, Ş. and Poland, J., eds., *Extraordinary Science and Psychiatry: Responses to the Crisis in Mental Health Research*, Cambridge, MA: MIT Press.
A critical evaluation of the crisis in mental health research following the publication of the *DSM*-5 through a Kuhnian conception of scientific change.

Thornton, T. (2007), *Essential Philosophy of Psychiatry*, Oxford: Oxford University Press.
Introductory overview.

Tsou, J. Y. (2012), Intervention, Causal Reasoning, and the Neurobiology of Mental Disorders: Pharmacological Drugs as Experimental Instruments, *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(2):542–551.
Argues that basic science and knowledge of the effects of medications can help scientists to understand the neurobiology of mental disorders.

Tsou, J. Y. (2016), Natural Kinds, Psychiatric Classification and the History of the DSM, *History of Psychiatry*, 27(4):406–424.
An evaluation of the evolution of the *DSM* under the light of the natural kind debate.

Wakefield, J. (1992a), The Concept of Mental Disorder: On the Boundary between Biological Facts and Social Values, *American Psychologist*, 47:373–388.
Presentation of the harmful dysfunction account of mental disorder.

Wakefield, J. C. (1992b), Disorder as Harmful Dysfunction: A Conceptual Critique of DSM-III-R's Definition of Mental Disorder, *Psychological Review*, 99(2):232–247.
A conceptual criticism of the *DSM* description of mental disorder by construing mental disorder as a harmful dysfunction.

Young, A. (1995), *The Harmony of Illusions: Inventing Post-Traumatic Stress Disorder*, Princeton, NJ: Princeton University Press.
A historical and ethnographic account of the evolution of post-traumatic stress disorder (PTSD) from the treatment of Vietnam veterans to contemporary scientific and clinical approaches to PTSD.

Zachar, P. (2014), *A Metaphysics of Psychopathology*, Cambridge, MA: MIT Press.
Develops a pragmatic approach to classifying mental disorders.

Index

addiction 230–1, 233–6, 271, 381, 433, 434, 436, 438–40, 445, 446, 546, 551
agency 9, 89, 103, 111, 154, 165, 223, 258–60, 264–8, 270–4, 278, 279, 309, 312, 317, 323, 325, 326–9, 331–9, 341–5, 350–5, 358, 364, 379, 382, 438, 443–6, 486, 488, 489, 498, 530, 536, 538, 540
ambivalence 182, 196, 441, 442
anorexia nervosa 14, 113–15, 117, 119, 121, 123–7, 129, 545
anti-realism 467–71, 473, 475–9, 481, 483
autism 14, 81, 93–5, 97, 99, 101, 103–9, 111, 155, 170, 200, 371–3, 375, 378–84, 386–9, 473–5, 482
autonomy 9, 13, 15, 16, 182, 198, 202, 213, 225, 240, 281, 284, 294–7, 301, 312, 333, 342, 344, 367, 382, 389, 433, 434–46, 530, 532, 537, 543, 544

basic emotions 93–9, 101–3, 108, 110, 111
Beauvoir, Simone De 133, 134, 152, 154
behaviorism 43, 45
biopower 15, 391–4, 396–9, 407
biopsychosocial model 14, 21, 28, 35, 41, 47, 53, 57, 58, 328, 546, 547
bipolar disorder 209, 211, 214, 330, 331, 395, 522, 545, 548
blame 15, 296, 314, 325–8, 331–7, 341, 343, 345, 347, 354, 355, 359–65, 443, 445
Boorse, Christopher 197, 213, 387, 451, 452, 463, 494, 499, 533, 534, 540, 543
borderline personality 81, 314, 344, 378, 455, 551

causality 52, 234–6, 448, 453–5, 457
causation
  absence and omission 221
  causal completeness 226
  counterfactuals 223
  dependency *vs* production 220–2
  interventionism 222–3
  locality 221
  mechanisms 222
  in medicine 218–20
  in science 218
  statistical dependence 222–3
  structural equations 223
  theories of 220
chemical imbalance 59, 63, 83, 87, 89, 90, 353, 364, 456
clinical case 16, 27, 36, 332, 543
clinical psychology 36, 193, 194, 234, 237–40, 244–8, 344, 448, 463, 544
coercion 15, 257, 268, 271, 283–93, 295–304, 312, 392
cognitive science 45, 50, 89, 117, 127, 128
compulsion 271, 284, 287, 302, 304, 436, 438, 441, 446
constructivism 93, 107, 108
crime predictive technology 391, 402
criminality 396, 398–404, 409, 410
critical psychiatry 521, 544

decision vectors 496
depression 7, 10, 14, 17, 20, 23, 29, 30, 43, 44, 47, 50, 52–4, 57, 62–5, 69, 70, 81–4, 86, 88–91, 104, 111, 124, 142, 143, 149, 150, 151, 153, 155–8, 161–6, 168, 170–2, 180, 190, 196, 204, 215, 221, 228, 248–51, 271, 273, 305, 314, 331, 340, 343, 345, 433, 434, 438, 450, 455–62, 464, 473, 478, 516, 524, 528–30, 535, 540, 543, 544, 547–9, 551
Descartes 42, 43, 47, 57
diagnosis 7, 19, 20, 24, 27, 28, 31–3, 57, 104–5, 181, 182, 193, 194, 200, 201, 209, 215, 320, 378, 389, 412, 416, 418, 422, 467, 469, 470, 471, 475, 477, 479–83, 488, 508–25, 552

disability 27, 31, 60, 109, 269, 319, 322, 327, 371, 375, 376, 379, 381, 387, 389, 410, 512, 528, 540
disciplinary power 391–4, 396, 397, 399, 402, 407
disease 7, 10, 11, 19–21, 23, 27, 28, 31, 32, 34, 41, 47–9, 72, 73, 75–7, 80, 88, 91, 114, 117, 149, 162, 180, 185, 186, 188, 206, 214, 217–19, 230, 235, 236, 275, 323, 347, 361–5, 380, 381, 394, 395, 438, 440, 446, 449–53, 462–4, 467, 488, 493, 494, 500, 501, 509, 511, 512, 514–16, 519, 521, 523, 525, 527, 528, 532–4, 537–40, 543, 544, 548, 551
DSM 7, 12, 15, 20, 27, 32, 33, 35, 43, 44, 51, 52, 80, 86, 110, 113, 114, 122, 124, 149, 161, 177, 178–96, 200, 202, 204–6, 208, 210, 212–15, 231, 233, 238, 239, 251, 317, 325, 327, 328, 330, 331, 338–42, 379, 383, 394, 412, 441, 464–5, 467, 475, 482, 483, 486, 494, 496, 497, 499, 500, 501, 502, 503, 511, 517, 519–21, 525, 540, 545, 546, 550–3
DSM-5 12, 20, 27, 35, 43, 51, 113, 114, 122, 124, 161, 185, 186, 194, 196, 200, 202, 204, 212, 214, 238, 239, 251, 325, 327, 328, 330, 331, 338, 339–42, 379, 383, 412, 464, 482, 486, 497, 500, 502, 503, 511, 517, 519, 525, 540, 545, 546, 550, 551, 553
dualism 42, 43, 45, 47, 48, 50, 53, 67, 69, 73, 76, 78, 240
dysfunction 22, 32, 46, 48, 51, 72–5, 77, 83, 188, 234, 328, 330, 354, 358, 373, 374, 377, 386–8, 452, 453, 460, 465, 493, 494, 499, 501, 502, 534, 550, 553

embodied mind 14, 113–15, 117, 119–25, 127–9
embodiment 14, 118, 120, 121, 123, 124, 127, 153, 170, 425
emotion 91, 94–100, 102–4, 106–11, 119–21, 126, 127, 188, 327, 336, 337, 365, 375, 382, 548
empathy 103, 105, 109, 111, 153, 156, 167, 168, 170–2, 352, 365, 388, 421

Engel, George 21, 47
epistemic agents 259, 260, 264, 268
epistemic realism 469, 476, 477, 481
epistemic values 198–207, 210–12, 215
epoché 135, 145–7, 150, 151
essentialist mechanisms 474
etiology 218–19, 228
evidence-based medicine 192, 219, 320, 495
existentialism 133, 153, 154
expertise 23, 33, 34, 202, 206, 210, 215, 259, 260, 486, 487, 489, 503, 539

Fanon, Frantz 15, 134, 415, 428
forced medication 284, 288, 289, 299
Foucault, Michel 15, 180, 193, 333, 391, 392–6, 402–4, 409–11, 539, 544, 546
free will 13, 15, 16, 37, 87, 129, 343–5, 433–7, 439–45, 530
functionalism 45, 46, 57
fusion of horizons 159, 171, 172, 421

Gadamer 14, 152, 155–61, 165–72, 421, 428
gender 9, 89, 98, 244, 252, 291, 310, 311, 316, 323, 374, 378, 395, 400–2, 408, 409, 467, 502, 542, 543, 551

Heidegger, Martin 125, 133, 142–4, 152, 153, 159
Hempel, C 183–5, 193, 199, 214
homeostatic mechanisms 472, 474–6
Hume, David 94, 111, 453–5, 457, 459, 464, 486, 501, 518, 523, 531
Husserl, Edmund 125, 133, 135–7, 142–8, 151–4
hysteria 396, 399, 401, 402, 411, 468

identity theory 14, 37, 44, 52, 58, 129
inductive risk 199, 203–5, 208, 213–15, 494, 500
intention 82, 118, 156, 168, 182, 224, 251, 261, 285, 294, 306, 314, 318, 388, 389, 416, 422, 437, 442, 515
intersubjectivity 6, 14, 133, 155–61, 163–8, 170, 171
interventions
  neural 227–8
  psychological 227–8
  social 227–8
  theory of causation 222–3

Jaspers, Karl 134, 151, 154, 549

levels of reality 224–7
lobotomy 396, 399–402, 408, 411
Locke, John 420, 459
looping effects 188–90, 193, 195, 548

major depressive disorder 20, 36, 51, 53, 81, 82, 88, 90, 91, 142, 161, 200–2, 204, 210, 211, 340, 452, 459, 461–2, 516, 540
manipulation 217, 232, 283, 284, 287, 293, 299, 312, 446
mental difference 69, 71, 72, 75, 82, 83
mental illness 4, 8, 10, 11, 17, 22, 37, 48, 56, 59, 109, 129, 178, 180, 194, 270, 271, 276, 305, 306, 307, 309, 311–25, 330–45, 347–64, 396–8, 401, 402, 434–46, 449–53, 465, 488, 493, 517–19, 523, 525, 527–41, 544, 548, 551–2
Merleau-Ponty, Maurice 14, 125, 133–54, 550
metaphysical realism 469, 470
mind-body problem 14, 41, 42, 46, 57, 217, 224
moral agency 111, 258, 260, 266–70, 271–9, 338, 350, 351, 355, 364, 379
moral responsibility 197, 240, 326, 343–5, 347–58, 360, 364 -366, 425, 442, 446
multiple realization 224–7, 228–32

natural function 451, 452
natural kinds 177, 178, 185, 187–92, 194–6, 470–3, 480, 482, 483, 504, 524, 544, 547–9, 553
neurodivergence 81, 374, 377
neurodiversity 11, 15, 60, 80, 81, 88, 109, 111, 371–89
neurodiversity paradigm 372–5, 377, 379–84, 386–8
neuroethics 15, 89, 235, 347–9, 351, 353–7, 359, 361, 363, 365, 367, 550
neuroscience 6, 8, 12, 36, 44, 46, 57, 63, 64, 71, 74–5, 77–80, 82, 83, 89–91, 111, 117, 126, 147, 170, 182, 186, 192–5, 214, 233–4, 240, 348–9, 354–5, 357, 364, 365, 428, 444, 446, 478, 479, 544, 545, 552
nominalism 459
nonepistemic values 197–212
North African Syndrome 15, 413, 415, 417, 419, 421, 423, 425–9
nudges 283, 284, 296, 300, 303

objective attitude 328, 350

pathology paradigm 373, 374, 378, 382, 386, 387, 389
persuasion 283, 284, 287, 291, 293, 294
philosophy of psychiatry 3, 4, 8, 9, 12, 13, 16, 552
physicalism 67–72, 76, 85, 86, 224–7
pluralism 8, 54, 213, 215, 308, 456, 464, 489, 495, 497, 523
Popper, C 447, 448, 464
Post Traumatic Stress Disorder (PTSD) 214, 238, 239, 247–52, 553
praise 325–8, 332–7
prognosis 22, 184, 294, 509
psychiatric classification 13, 15, 16, 177–89, 191, 193–7, 199, 201, 203, 205–9, 211, 213, 215, 391, 397–9, 409, 447–9, 451, 453, 455, 457, 459, 461, 463, 465, 483, 486, 489, 490, 497, 498, 520, 524, 525, 553
psychopathy 14, 93–5, 97, 99, 101, 103–11, 339, 340, 347, 348, 363, 364, 366, 550
psychotherapy 22–4, 28, 30, 35, 54, 78, 88, 90, 180, 227–8, 239, 248, 251, 344, 345, 458, 500, 530, 545

race 244, 310, 311, 315, 316, 322, 323, 392–5, 407–11, 418, 419, 424, 427, 464, 542, 544
reactive attitudes 327, 328, 332–40, 349–54, 359–65, 443
relativism 54, 96, 102, 104
Research Domain Criteria (RDoC) 12, 44, 57, 91, 182, 183, 186, 187, 192–4, 231, 233–5, 545, 552
research ethics 276, 283–6
restraint 27, 284, 287–9, 301
rubber hand illusion 121

schizophrenia 10, 17, 29, 62, 81, 123, 134, 140, 142, 143, 151, 155, 156, 170, 171, 180, 188, 194, 230, 304, 314, 323, 326, 331, 341, 343, 371–5, 379, 383–8,

433, 438, 449–51, 460, 461, 465, 468, 476, 478, 479, 483, 494, 499, 516, 522, 528, 531, 544, 545, 548, 552
scientific realism 467–9, 471, 473, 475–7, 479, 481, 483
seclusion 288, 289, 301
sense of self 117, 118, 120–2, 124, 143, 242, 309, 313, 384
sentimentalism 93, 110
social empiricism 496, 503
social epistemology 16, 485–95, 497–501, 503
social model of disability 375
social objectivity 495, 496
Steel, Daniel 198, 199, 203, 205, 215
sterilization 15, 391, 396–9, 408, 410, 411
Szasz, Thomas 10, 11, 17, 180, 195, 449–51, 465, 515, 517–19, 525, 528–32, 535–541, 544, 552

testimonial justice 487, 488
therapeutic alliance 271, 337, 338, 413
trauma 15, 87, 237–53, 266, 267, 309, 324, 401, 402, 411, 548
trauma-informed 15, 237, 239–45, 247–53
type/token distinction 14, 55, 56

underdetermination 203, 204, 207, 213, 478, 479, 491, 498, 502

value-free ideal 197, 202–9, 212, 213
value-laden ideal 197, 198, 203, 492, 496

Wakefield, Jeremy 51, 52, 57, 74, 91, 197, 215, 387, 388, 452, 453, 465, 494, 499, 501, 503, 511, 525, 534, 540, 541, 548, 550, 553
well-ordered Science 492, 493